USING SPARSE CODING FOR LANDMARK LOCALIZATION IN FACIAL EXPRESSIONS

Vittorio Cuculo, Raffaella Lanzarotti*, Giuseppe Boccignone

Dipartimento di Informatica, Università di Milano
Via Comelico 39/40, 20135 Milano, Italy
vittorio.cuculo@studenti.unimi.it
lanzarotti,boccignone@di.unimi.it

ABSTRACT
In this article we address the issue of adopting a local sparse coding representation (Histogram of Sparse Codes), in a part-based framework for inferring the locations of facial landmarks. The rationale behind this approach is that unsupervised learning of sparse code dictionaries from face data can be an effective approach to cope with such a challenging problem. Results obtained on the CMU Multi-PIE Face dataset are presented providing support for this approach.

Index Terms— Facial landmarks, sparse coding, part-based models

1. INTRODUCTION
Facial expression, together with other communicative signals such as head gestures and gaze, is a rich source of information for nonverbal interaction that humans ceaselessly exploit to coordinate their activities and social relationships [1, 2].

Thus, expression recognition has attracted much attention in the emerging field of Social Signal Processing (SSP, [3]) and it has been widely investigated in the past decade, but it remains an open research issue due to the complexity and variety of expressions. Unless an holistic approach to recognition is pursued, where the face is represented as a whole unit [4, 5], then a crucial initial step is the detection of certain facial feature points usually referred as landmarks (albeit the holistic / local distinction is somehow elusive, if the holistic view is simply advocated to state that a facial part is not perceived independently of the other parts, cfr. [6]).

A face landmark can be defined [7, 8] as a prominent feature that can play a discriminative role or can serve as anchor point on a face graph: for example, the end points of the eyebrow arcs, the eye corners, the nose tip, the nostril corners, the mouth corners. Landmark detection is an interesting and yet unsolved problem, especially when performed "in the wild" [7]. In this case, one is likely to deal with variability in pose, lighting, age, gender, race, make-up, etc.

At the most general level, the landmark localisation problem can be summarised as follows. Let \( \mathbf{L} = \{l_1^1, l_2^1, \ldots, l_n^1\} \) denote the locations of \( n \) landmarking parts of the face. Let \( \mathbf{F} = \{f_1^1, f_2^1, \ldots, f_n^1\} \) denote the measured detector responses, where \( f_i^1 = \phi(l_i^1, I) \) is the response or feature vector provided by a local detector at location \( l_i^1 \) in image or frame \( I \). Then, localisation can be solved by finding the value of \( \mathbf{L} \) that maximises the probability of \( \mathbf{L} \) given the responses from local detectors, namely \( \mathbf{L}^* = \arg \max_{\mathbf{L}} P(\mathbf{L}|\mathbf{F}) \).

In the work presented here we learn from facial images an effective local representation by resorting to the sparse coding framework. In particular, we adopt the Histograms of Sparse Codes (HSC) that have been proposed in [9] for generic object detection, but have been never been experimented for facial landmark localization.

The paper is organised as follows. In Section 2 we provide some background and motivations for the issue of learning reliable landmark detectors. In Section 3 we discuss the methodological framework adopted in our approach, namely the use of sparse coding learning of landmarks within a probabilistic part-based detection approach. Experiments and dataset used are presented and discussed in Section 4, while some conclusions are drawn in Section 5.

2. BACKGROUND AND MOTIVATIONS
The spatial configuration and temporal dynamics of landmarks can provide a viable way to analyze facial expressions and to objectively describe head gestures and facial expressions. Selected landmarks can be exploited for automatic identification of Action Units (AUs, e.g.,[10]) within the framework of the Facial Action Coding System (FACS), which has been designed to detect independent subtle changes in facial appearance caused by contractions of the facial muscles [11]. Alternatively, landmarks can be exploited for expression classification without resorting to AU recognition and coding [12]. Beyond the field of SSP, landmarking for face coding, tracking and recognition has fostered a huge

---

*The research was carried out as part of the project "Interpreting emotions: a computational tool integrating facial expressions and biosignals based shape analysis and bayesian networks", which is supported by the Italian Government, managed by MIUR, financed by the Future in Research Fund.
literature (for an in-depth and recent review, see for instance [7]). For what specifically concerns the work presented here, we want to spotlight two trends that are crucial for landmark detection within the expression recognition perspective.

The first is the wider use of machine learning techniques, not only for final classification purposes but also at the feature selection and extraction stage, namely to derive the response ϕ(λ, T). Instead of using hand-crafted descriptors (SIFT, HoG, Gabor, LBP), an effective representation can be learnt directly from images in an unsupervised data-driven manner [13].

In this perspective, we adopt sparse coding that has been effectively used in a variety of tasks in computer vision, such as face recognition [14, 15] and object detection [16, 9], and it is appealing in the analysis of affective expression [17].

The second trend is the inference of landmarks depending on or together with other visual cues, head pose being the most remarkable one. This leads to approaches based on part-based descriptions, where one can learn descriptions of individual parts and then compose them, generalizing to an exponential number of combinations [18]. Interestingly enough, relatively few studies investigated the fusion of the information from facial expressions and head movement, e.g., [19, 20, 21].

3. METHODS

In a part-based detection approach every facial landmark can be modeled as a part, and it can be assumed that the locations L of parts of the face can be generated according to m views or poses by some similarity transformation t, giving rise to the global model L_{k,t} [22]. Then, the generation of L can be accomplished by marginalising over the set of m models, i.e.,

\[ P(L|F) = \sum_{k=1}^{m} \int P(L|L_{k,t}) P(L_{k,t}|F) dt. \] (1)

Here the term \( P(L|L_{k,t}) \) accounts for dependence of L from the global configuration L_{k,t}.

Let us assume, in the vein of [23], that: i) the locations of the parts \( L' \) are conditionally independent of one another and the same holds for the detector responses F; ii) the relation between the transformed model landmark and the true landmark is translationally invariant, i.e., \( P(L_{k,t}|L_{k,t}) \) only depends on \( \Delta L_{k,t} = L_{k,t} - L' \). Then, by using Eq. 1, the following MAP solution can be derived,

\[ L^* = \arg \max_{L} \sum_{k=1}^{m} \int \prod_{i=1}^{n} P(\Delta L_{k,i}|L_{k,t}) P(L'|F') dt, \] (2)

where the prior \( P(\Delta L_{k,i}) \) accounts for the shape or global component of the model, and \( P(L'|F') \) for the appearance or local component. For instance, \( P(L'|F') \propto \exp(\mathbf{u}_k \cdot \mathbf{F}') \), where the parameter \( \mathbf{u}_k \) is a template tuned for model k at location \( L' \).

There can be a variety of solutions to solve or approximate the problem posed in Eq. 2 [24]. For instance, Belhumeur et al. [23] straightforwardly address the part localization in Eq. 2 via Bayesian inference. In the experiments presented in Section 4 we exploit Zhu & Ramanan [25], which draws on the part-based model of Felzenszwalb et al. [18]. This solution relies upon a linearly-parameterized, tree-shaped pictorial structure of the landmark rich parts of the face [24]. Precisely, such a structure is based on mixture of trees with a shared pool of parts. Local and global information is merged from beginning via the tree-connected patches covering the landmarkable zones of the face (68 landmarks for frontal and 39 landmarks for profile faces). Consistently with the assumption of translational invariance, \( P(\Delta L_{k,i}) = N(\Delta L_{k,i}; \mu, \Sigma) \), i.e., it is a gaussian distribution, which can be interpreted as a spring connecting two parts, with rest position \( \mu \) and a stiffness \( \Sigma \). In such framework, Eq. 2 is solved via the optimized search of the maximum likelihood tree.

However, for what concerns the local component \( P(L'|F') \), [25] relies on patches representing HoG responses to face landmarks. As previously discussed, and more cogently in the case of face expression analysis, a more effective representation can be learnt directly from example images, rather than using hand-crafted descriptors. Thus, to sample patch responses \( F' \) we resort to the sparse coding approach [26], which has recently gained currency in face analysis [14, 17, 15].

Denote the matrix of observed image patches \( X = [x_1, \ldots, x_n] \in \mathbb{R}^{D \times N} \). Denote matrix \( W = [w_1, \ldots, w_p] \in \mathbb{R}^{D \times W} \) a dictionary and \( Z = [z_1, \ldots, z_n] \in \mathbb{R}^{W \times N} \) the associated sparse code matrix. Each column \( w_i \) is referred to as an atom. Each observed vector \( x_i \) is approximated as a sparse combination of basis vectors \( w_i \), i.e. \( x = Wz + v \), \( v \) being a residual noise vector sampled from a zero mean Gaussian distribution \( \mathcal{N}(v; 0, \sigma^2 I) \). Sparse coding is the process of computing the representation coefficients \( z \) based on the given signal \( x \) and the dictionary \( W \). Note that the sparsity pattern (controlled by \( z_i \)) changes from data case to data case. When \( W > D \) the representation is called overcomplete. It is possible to learn the dictionary, by maximizing the likelihood

\[ \log P(X|W) = \sum_{i=1}^{N} \log(\mathcal{N}(x_i|Wz_i, \sigma^2 I)) + \log P(z_i). \] (3)

The maximization of the log-likelihood in Eq. 3, can be turned in the minimization of the negative log-likelihood (NLL). This can be done efficiently by using the K-SVD algorithm [28]. The K-SVD minimizes the NLL via the
objective function

$$\min_{W,Z} \|X - WZ\|^2 \text{ subject to } \forall i, \|x_i\|_0 \leq K,$$  

where the zero-norm \(\|\cdot\|_0\) counts the non-zero entries in the sparse code \(z_i\), and \(K\) is a pre-defined sparsity level.

Optimization is solved by alternating between computing \(Z\) (sparse coding stage via Orthogonal Matching Pursuit, OMP [28]) and \(W\) (codebook update stage by Singular Value Decomposition). Interestingly enough, the dictionary that is learned by applying sparse coding to patches of natural images consists of basis vectors that look like the filters that are found in simple cells in the primary visual cortex of the mammalian brain [27].

Once the dictionary \(W\) is learned, it is possible to use OMP to compute sparse codes at every location in an image. Here, in the vein of [9], we adopt the Histograms of Sparse Codes (HSC) representation to sample the local response \(f_i\). Briefly, HSC is a (semi-)dense feature vector \(F\) obtained averaging the sparse codes of patches in a \(16 \times 16\) neighborhood (see [9] for details). The final response \(f_i\) is obtained normalizing \(F\) and augmenting its discriminative power applying the transform \(F^{\mathbf{e}}\). The HSC maintains the local response aggregation similarly to HoG, but with the advantage of learning richer features on patches by capturing more information than gradients, while avoiding ad-hoc design choices.

To sum up, the framework introduced here can cope with both issues discussed in Section 2: local feature learning and joint landmark and pose inference.

In this article we are primarily concerned with the assessment of HSC performance for landmarking, thus, in the following Section we will discuss results obtained on frontal faces with different expressions and acquired under different illumination conditions, and compare them with the ones obtained adopting the HoG feature.

4. EXPERIMENTAL RESULTS

We refer to a subset of the frontal face images of the CMU Multi-PIE Face database, namely the one corresponding to the annotations made available by Torr et al. [29] (66 points on 734 frontal face images). Such set has been partitioned in three parts: the first 250 for the model learning, the following 250 for the dictionary learning, and the last 234 for the testing phase. In addition, for the model learning phase only, we use the whole set of 1218 non-face images from the INRIA Person dataset [30].

The dictionary learning has been arranged determining at first the best setup on patches of \(5 \times 5\) dimensions, thus varying the sparsity level \(k\) between 1 and 3. The best level achieved has then been adopted for a further investigation on patches of dimension \(7 \times 7\) (i.e. with \(k = 1\)).

The landmark localization error has been evaluated using the RMS error normalized with respect to the face size, that is the mean of its height and width. In Fig. 1 we plot the obtained results. The best results are attained by using the \(5 \times 5\) HSC feature with \(k = 1\). Indeed, the 97\% of the images tested with this descriptor has revealed a localization error lower than the 5\% of the face size, contrary to the 87.55\% obtained using the HoG descriptor. We also randomly split the test images in 10 subsets and extract the landmarks adopting both the HoG and the HSC features. This leads to a mean localization error of 0.05 and 0.03 respectively, with very low standard deviations in both cases (8.5e − 06 and 3.7e − 06). Considering the HoG descriptor has demonstrated in [25] to be the state-of-the-art in this field (it overcomes [31] by 5.7\% [32] of 8.8\%, [33] of 9.7\%, and [34] of 20.4\%), these results highlight the effectiveness of the HSC-based method for landmark localization.

In Fig. 2 and Fig. 3 we report some examples of the algorithm behavior when applied to non neutral face images and to realistic face videos respectively. As we can visually appreciate, in both cases the landmarks describe faithfully the facial feature locations, even in presence of closed eyes, open mouth, and beard. Finally, even if out of the scope of this article, in Fig. 4 we show the behavior of the method when applied to non frontal faces. These preliminary results encourage the application of the method to non frontal images, while requiring a proper dictionary learning to attain higher precision.

Concerning the computational time, the method based on the \(5 \times 5\) HSC results more expensive than the HoG-based: they require 18.30 and 8.75 seconds respectively to make an inference on a face image of \(640 \times 680\) pixels. Such compu-

---

1This computational times have been obtained using a Quad Core i7-
tational costs could be significantly reduced parallelizing the code.

5. CONCLUSIONS

In this article we have addressed the issue of integrating the HSC technique in a part-based framework suitable to provide the joint inference of facial landmarks and pose. Results obtained so far substantiate that learned sparse codes can be more expressive and effective than gradients for facial landmarking, only involving a moderate computational cost. HSC can outperform the widely use HoG representation, while avoiding some well-known pitfalls of HoG tuning. Yet, the subtleties and specificities brought out by facial features as opposed to general object detection (e.g., birds, boats, cars, etc. [9]) should not be overlooked, one example being the choice of the sparsity level.

The promising results obtained on frontal faces encourage us to extend this approach to multi-pose face images, thus making the system suitable to deal with real applications such as emotion or face recognition.

6. REFERENCES


[3] A. Vinciarelli, M. Pantic, D. Heylen, C. Pelachaud,


[34] http://face.com/.


