



Deep fusion of visual signatures for client-server facial analysis

Binod Bhattarai, Gaurav Sharma, Frédéric Jurie

► To cite this version:

Binod Bhattarai, Gaurav Sharma, Frédéric Jurie. Deep fusion of visual signatures for client-server facial analysis. Tenth Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP 2016), Dec 2016, Guwahati, India. hal-01390001v2

HAL Id: hal-01390001

<https://hal.science/hal-01390001v2>

Submitted on 9 Nov 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Deep fusion of visual signatures for client-server facial analysis

Binod Bhattarai^{*}
Normandie Univ, UNICAEN,
ENSICAEN, CNRS, GREYC
binod.bhattarai@unicaen.fr

Gaurav Sharma[†]
Computer Sc. & Engg.
IIT Kanpur, India
grv@cse.iitk.ac.in

Frederic Jurie[‡]
Normandie Univ, UNICAEN,
ENSICAEN, CNRS, GREYC
frederic.jurie@unicaen.fr

Facial analysis is a key technology for enabling human-machine interaction. In this context, we present a client-server framework, where a client transmits the signature of a face to be analyzed to the server, and, in return, the server sends back various information describing the face e.g. is the person male or female, is she/he bald, does he have a mustache, etc. We assume that a client can compute one (or a combination) of visual features; from very simple and efficient features, like Local Binary Patterns, to more complex and computationally heavy, like Fisher Vectors and CNN based, depending on the computing resources available. The challenge addressed in this paper is to design a common universal representation such that a single merged signature is transmitted to the server, whatever be the type and number of features computed by the client, ensuring nonetheless an optimal performance. Our solution is based on learning of a common optimal subspace for aligning the different face features and merging them into a universal signature. We have validated the proposed method on the challenging CelebA dataset, on which our method outperforms existing state-of-art methods when rich representation is available at test time, while giving competitive performance when only simple signatures (like LBP) are available at test time due to resource constraints on the client.

1. INTRODUCTION

We propose a novel method in a heterogeneous server-client framework for the challenging and important task of analyzing images of faces. Facial analysis is a key ingredient for assistive computer vision and human-machine interaction methods, and systems and incorporating high-performing methods in daily life devices is a challenging task. The objective of the present paper is to develop state-of-the-art technologies for recognizing facial expressions and facial attributes on mobile and low cost devices. Depending on their computing resources, the clients (i.e. the devices on which the face image is taken) are capable of computing different types of face signatures, from the simplest ones (e.g. LBP)

to the most complex ones (e.g. very deep CNN features), and should be able to eventually combine them into a single rich signature. Moreover, it is convenient if the face analyzer, which might require significant computing resources, is implemented on a server receiving face signatures and computing facial expressions and attributes from these signatures. Keeping the computation of the signatures on the client is safer in terms of privacy, as the original images are not transmitted, and keeping the analysis part on the server is also beneficial for easy model upgrades in the future. To limit the transmission costs, the signatures have to be made as compact as possible. In summary, the technology needed for this scenario has to be able to merge the different available features – the number of features available at test time is not known in advance but is dependent on the computing resources available on the client – producing a unique rich and compact signature of the face, which can be transmitted and analyzed by a server. Ideally, we would like the universal signature to have the following properties: when all the features are available, we would like the performance of the signature to be better than the one of a system specifically optimized for any single type of feature. In addition, we would like to have reasonable performance when only one type of feature is available at test time.

For developing such a system, we propose a *hybrid deep neural network* and give a method to carefully fine-tune the network parameters while learning with all or a subset of features available. Thus, the proposed network can process a number of wide ranges of feature types such as hand-crafted LBP and FV, or even CNN features which are learned end-to-end.

While CNNs have been quite successful in computer vision [1], representing images with CNN features is relatively time consuming, much more than some simple hand-crafted features such as LBP. Thus, the use of CNN in real-time applications is still not feasible. In addition, the use of robust hand-crafted features such as FV in hybrid architectures can give performance comparable to Deep CNN features [2]. The main advantage of learning hybrid architectures is to avoid having large numbers of convolutional and pooling layers. Again from [2], we can also observe that hybrid architectures improve the performance of hand-crafted features e.g. FVs. Therefore, hybrid architectures are useful for the cases where only hand-crafted features, and not the original images, are available during training and testing time. This scenario is useful when it is not possible to share training images due to copyright or privacy issues.

Hybrid networks are particularly adapted to our client-

server setting. The client may send image descriptors either in the form of some hand-crafted features or CNN features or all of them, depending on the available computing power. The server has to make correct predictions with any number and combination of features from the client. The naive solution would be to train classification model for the type of features as well as for any of their combinations and place them in the server. This will increase the number of model parameters exponentially with the number of different feature types. The proposed hybrid network aligns the different feature before fusing them in a unique signature.

The main contribution of the paper is a novel multi-features fusion hybrid deep network, which can accept a number of wide ranges of feature types and fuse them in an optimal way. The proposed network first processes the different features with feature specific layers which are then followed by layers shared by all feature types. The former layer(s) generate(s) compact and discriminative signatures while the later ones process the signatures to make predictions for the faces. We learn both feature specific parameters and shared parameters to minimize the loss function using back propagation in such a way that all the component features are aligned in a shared discriminative subspace. During test time, even if all the features are not available, e.g. due to computation limitations, the network can make good predictions with graceful degradation depending on the number of features missing.

The thorough experimental validation provided, demonstrates that the proposed architecture gives state-of-the art result on attributes prediction on the CelabA dataset when all the features are available. The method also performs competitively when the number of features available is less i.e. in a resource-constrained situation.

The rest of the paper is organized as follows: Sec. 2 presents the related works, Sec. 3 gives the details of our approach while Sec. 4 presents the experimental validation.

2. RELATED WORKS

In this section we review some of the works which are, on one side, related to hybrid architectures or, on the other side, related to multimodal fusion and face attribute classification. Apart from face attributes classification, other critical applications on faces are: large scale face retrieval [3, 4], face verification [5, 6, 7, 8], age estimation [9, 10], etc. For more details on the application of faces and comprehensive comparison of recent works, we suggest the readers refer [11].

Hybrid Architectures. One of the closest works to our work is from Perronnin et al. [2]. The main idea behind their work is to use Fisher Vectors as input to Neural Networks (NN) having few fully connected (supervised) layers (up to 3) and to learn the parameters of these layers to minimize the loss function. The parameters are optimized using back propagation. Unlike their architecture, our network takes a number of wide range of hand-crafted features including FVs, but not only. In addition, our architecture is also equipped with both feature specific parameters and common parameters. We have designed our network in such a way that the input features are aligned to each other in their sub-spaces. The advantage of such alignments is that our system can give good performance even when a single type of feature is present at test time. Moreover, such ability

makes our system feature independent i.e. it can properly handle any types of features it encounters.

There are some works, such as [12], which, instead of taking hand-crafted features as input, takes CNN features and compute FVs in the context of efficient image retrieval and image tagging. This approach improves the performance of CNNs and attains state-of-art performance, showing that not only FVs but also CNNs benefit from hybrid architecture.

Face Attribute Classification. Some of the earliest and seminal work on facial attribute classification is the works from Kumar et al. [13, 14]. Both of their papers use hand-crafted low-level features to represent faces, sampled with AdaBoost in order to discover the most discriminative ones for a given attribute, and train binary SVM classifiers on this subset of features to perform attribute classification. The current state-of-art method of Liu et al. [15] uses two deep networks, one for face localization and another for identity based face classification. The penultimate layer of the identity classification network is taken as the face representation, and a binary SVM classifier is trained to perform an attribute classification. Some other recent state-of-the-art methods such as PANDA [16], Gated ConvNet [17], etc. also use deep learning to learn the image representation and do attribute classifications on it. From these works, we can observe that either hand-crafted features or CNN features are used for attribute classification. From our knowledge, the proposed method is the first to learn a hybrid structure combining multiple hand-crafted and CNN features for facial attribute classification. Moreover, most of the mentioned works here are performing binary attribute classification while we are predicting multiple attributes of faces.

Multi-modal fusion. Recently Neverova et al. [18] proposed a method called *Mod-Drop* to fuse information from multiple sources. Their main idea is to take a batch of examples from one source at a time and feed into the network to learn the parameters, instead of taking examples from all the sources. The main drawbacks of their approach is, when a new source is encountered and is to be fused, it requires to re-train the whole network. Some other recent works such as [19, 20, 21, 22] fuse multiple sources of information to improve the performance of the final result. None of these works evaluated the performance of component sources or their possible combinations after fusion.

3. APPROACH

As mentioned before, a key challenge addressed in this paper is to learn an optimal way to fuse several image features into a common signature, through the use of a hybrid fully connected deep network. This section presents the proposed method in detail, explains how to learn the parameters and gives technical details regarding the architecture.

3.1 Network architecture

Fig. 2 shows a schematic diagram of the proposed network. A, B and C denote the different feature types to be aligned and fused, which are the input to the network. We recall that all or only a subset of the features can be available depending on the computing resources of the client. While we show a network with 3 features types, more can be used with similar layers for the new features. The key idea here

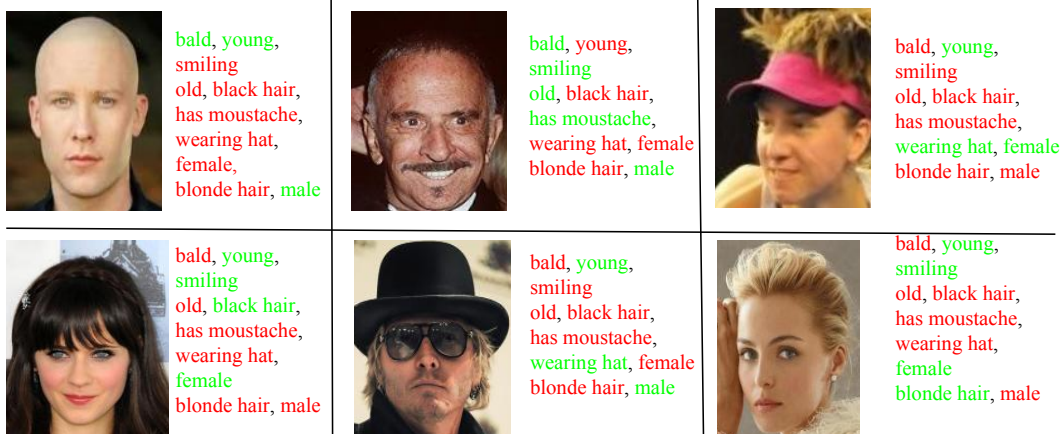


Figure 1: Randomly sampled images of CelebA and a subset of attributes. Green color attributes are relevant for the image whereas red color attributes are irrelevant (better viewed in color).

is to train a single network which consists of feature specific layers (shown in blue), to be implanted on the clients, and common layers (shown in black), to be implanted on the server. The activation of the middle layer, obtained after merging the feature specific layers, gives the universal signature which will be transmitted from the client to the server. Each layer is fully connected with its parents in the network. In our application the output of the network is the facial expressions/attributes to be recognized, one neuron per expression/attribute, with the final values indicating the score for the presence of these attributes.

3.2 Learning the parameters of the network

Carefully setting up the learning of such hybrid network is the main issue for competitive performance. We propose to learn the parameters of this network with a multistage approach. We start by learning an initialization of the common parameters. To do this we work with the most discriminate feature type (e.g. A, B or C). For example, suppose we observed that A is the most discriminate for our application (as discussed in the experiment section, we will see that for our application FVs are the most discriminant features). Thus we start learning the parameters of the network corresponding to both (i) the feature specific parameters of network A (blue layers) and (ii) the part of the network common to all features (black layers). Then we fix the common parameters and learn the feature specific parameters of the feature B taking training examples encoded with B. In our case, the task is same but the features are different during each training round. By repeating the same procedure, we learn the feature specific parameters of the network for each of the remaining type of features. In the end, all the features are aligned into a common signature which can then be transmitted to the server for the computation.

The major advantage of this strategy is that although we are mapping all the features into same feature space, we do not require feature to feature correspondence e.g. we are not using a certain feature type to estimate or mimic any other feature type. Moreover, when we encounter a new feature type, we can easily branch out the existing network and learn its parameter without hindering the performance of other feature types. Thus the proposed learning strategy, while performing very well, also avoids the retraining of the whole network upon addition of a new features type.

This is a major advantage of this our approach over existing Mod-drop [18] algorithm. Finally, since there are fewer parameters to optimize than training one distinct network per feature, the computations required are less and the training is faster.

Another alternative, that we explored, is to learn the parameters of the whole network first with all the available feature types, and then fix the common parameters and fine-tune the feature specific parameters. The reason behind this approach is to make shared subspace more discriminative than with the one learned with the single most discriminative feature so that we can align all the component features in this subspace and improve the overall performance. We found the performance obtained with this approach is slightly better than the one we discussed before. However, this alternative requires feature to feature correspondence mapping. Moreover, training with all the features at a time requires more computing resource and also leads to slow convergence and longer training time. We compare the performances of these methods in more details in the experiment section.

3.3 Details of the architecture

The proposed network is composed of only fully connected (FC) layers. Once the features are fed into the network, they undergo feature specific linear projections followed by processing with Rectified Linear Units (ReLU). Eq. 1 gives the feature-specific transformations, where σ is the non-linear transformation function i.e. ReLU, W_A, W_B, W_C and $\mathbf{b}_A, \mathbf{b}_B, \mathbf{b}_C$ are projection matrices and biases for the input features of the networks A, B, and C respectively. These representations further go into linear projections followed by ReLU depending upon the depth of the network.

$$\begin{aligned} h^A &= \sigma(\mathbf{x}_A W_A + \mathbf{b}_A) \\ h^B &= \sigma(\mathbf{x}_B W_B + \mathbf{b}_B) \\ h^C &= \sigma(\mathbf{x}_C W_C + \mathbf{b}_C) \end{aligned} \quad (1)$$

When the network takes more than one type of features at a time, it first transforms them with the FC and ReLU layers and then sums them and feeds into the common part of the network. We call this step as *merging*, as shown in the diagram. We further call the vector obtained at this point, after merging, as the signature of the face.

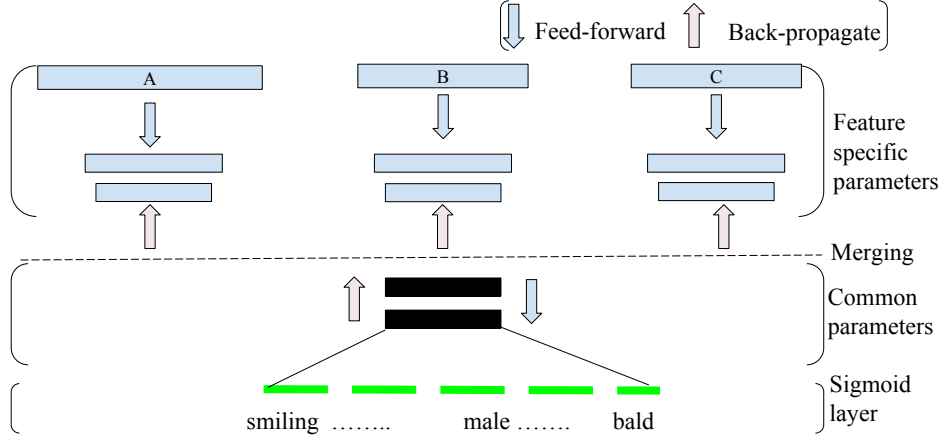


Figure 2: Illustration of proposed method.

Parameters Type	Layer Type	A	B	C
Feature Specific	Input	\mathbf{x}_A	\mathbf{x}_B	\mathbf{x}_C
	FC(ReLU)	4096	4096	4096
	FC(ReLU)	1024	1024	1024
Merge	Add	1024		
Common	FC(ReLU)	1024		
	FC(ReLU)	1024		
	Sigmoid	40		

Table 1: Details of parameters of proposed network

In the common part of the network, intermediate hidden layers are projected into linear space followed by ReLU. The final layer of the network is a sigmoid layer. Since we are doing multilabel predictions, sigmoid will assign higher probabilities to the ground truth classes. We learn the parameters to minimize the sum of binary cross-entropy of all the predictions of the sigmoid layer. We minimize the loss function using Stochastic Gradient Descent (SGD) with standard back propagation method for network training.

In the heterogeneous client-server setting, the client is expected to compute the signature and send it to the server for processing. Since different clients can have very different computing capabilities they can compute their signature with different types and number of features – in the worst case with just one feature. The method allows for such diversity among clients and as the server side works with the provided signature while being agnostic about what and how many features were used to make it.

4. EXPERIMENTS

We now present the experimental validation of the proposed method on the task of facial attribute classification. All the quantitative evaluation is done on the CelebA dataset [15], the largest publicly available dataset annotated with facial attributes. There are more than 200,000 face images annotated with 40 facial attributes. This dataset is split into train, val, and test sets. We use train and val set for training and parameter selection respectively, and we report the results obtained on the test set.

In the rest of the section, we first give the implementation

details and then discuss the results we obtained.

4.1 Implementation details

We have performed all our experiments with the publicly available aligned and cropped version of the CelebA¹ [15] dataset (without any further pre-processing). We assume that up to 3 different types of features can be computed, namely, Local Binary Patterns, Fisher Vectors and Convolutional Neural Networks features, as described below.

Local Binary Patterns (LBP). We use the publicly available `vlfeat` [23] library to compute the LBP descriptors. The images are cropped to 218×178 pixels. We set cell size equal to 20, which yields a descriptor of dimension 4640.

Fisher Vectors (FV). We compute Fisher Vectors following Simoyan et al [6]. We compute dense SIFTs at multiple scales, and compress them to a dimension of 64 using Principal Component Analysis. We use a Gaussian mixture model with 256 Gaussian components. Thus, the dimension of the FV feature is of 32,768 ($2 \times 256 \times 64$). The performance of this descriptor is $77.6 \pm 1.2\%$ on LFW for the task of face verification, with unsupervised setting, which is comparable to the one reported [6].

Convolutional Neural Networks (CNN). We use the publicly available state-of-art CNN mode trained on millions of faces presented in [7], to compute the CNN features. The dimension of CNN feature is of 4096. Our implementation of this feature gives $94.5 \pm 1.1\%$ on LFW for verification in unsupervised setting. Here, these features are computed without flipping and/or multiples of cropping of faces.

4.2 Baseline methods.

We report two different types of baselines. In the first one, the network is trained with a given feature type (e.g. LBP) while the same type of feature is used at test time (e.g. LBP again). We call this type of network as *Dedicated Networks*. In the second setting, we allow the set of features at train time and the one used at test time to differ. Such networks are adapted to different sets of features. This is the particular situation we are interested in. More precisely, we

¹<http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html>

Method	Avg. Precision
Random	23.1%
FVNet	69.0%
CNNNet	68.7%
LBPNet	64.3%

Table 2: Average Precision (AP) of single feature type baselines

experimented with 3 different dedicated networks (one per feature type) and 2 adapted networks, as detailed below, all such are considered as baselines.

LBPNet/FVNet/CNNNet. These baseline networks use only LBP, FV or CNN features, respectively, for both training and testing. They provide the single feature performances, assuming that no other feature is available either at training or testing.

All Feature Training Network (AllFeatNet). In this setting, all the available features are used to train the network. At test time, one or more than one type of features can be used, depending on its availability. For us, the available features are as described before FVs, CNNs, and LBPs.

Mod-Drop. This is currently the best method for learning cross-modal architectures, inspired by [18]. It consists, at train time, in randomly sampling a batch of examples including only one type of features at a time, instead of taking all the available features, and learn the parameters in a stochastic manner. We refer the reader to the original work [18] for more details.

4.3 The proposed method.

On the basis of which we fix the parameters of the common shared subspace, we categorize the proposed methods into two:

FVNetInit. Tab. 2 shows the individual performance of different features we used for our experiments. From the table we can see that FVs are most discriminative for our application. Thus, we choose to take few top layer’s parameters (please refer Tab. 1 of for the number of layers in shared subspace) of FVNet as common shared parameters of proposed network. Once we fix this, we learn the feature specific parameters for CNNs and LBPs to minimize the loss function. Fig. 4 shows the evolution of performances of FVs, LBPs, and CNNs with the amount of training epochs.

AllFeatNetInit. In this case, we use the common part of AllFeatNet as a starting point. Then we fix these parameters and learn the feature specific parameters of FVs, LBPs and CNNs to minimize the loss the function.

4.4 Quantitative results

We now present the results of the experiments we do to evaluate the proposed method. We measure the performance using average precision (AP) i.e. the area under the precision vs. recall curve. We do not consider attribute label imbalances for all the cases, unless explicitly stated.

Our experiments are mainly focused on validating two as-

Method	mean Avg. Precision
AllFeatNet	$63.4 \pm 9.5 \%$
Mod-Drop	$67.8 \pm 3.7 \%$
Ours(FVNetInit)	$68.8 \pm 3.0\%$
Ours(AllFeatNetInit)	$69.0 \pm 3.4\%$

Table 3: mean AP(mAP) of multi-feature baselines

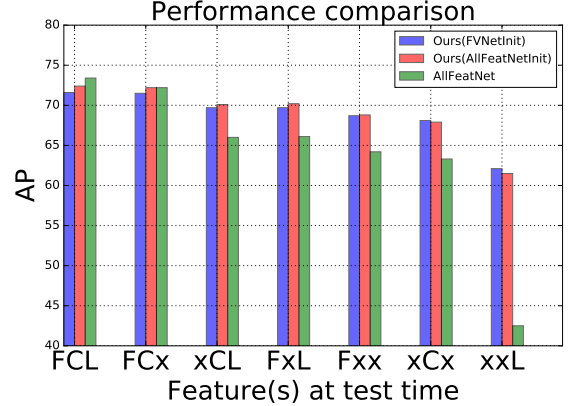


Figure 3: Performance comparison between different methods and different combinations of feature(s) at test time. FCL represents FVs, CNNs, and LBPs respectively. 'x' denotes the absence of the corresponding feature.

pects of the proposed method. First, we demonstrate that the performance due to individual features are retained after merging all the features in the same common subspace. Second, we demonstrate that the performance is improved in the presence of more information, i.e. presence of multiple types of features at a time.

Performance comparison with Dedicated Networks.

Tab. 2 and Tab. 4 give the performance of single features networks and their comparison with that of the multi-feature trained network (when, at test time, only one type of feature is present). From these tables, we can observe that, with both our approaches, the performance of the component features at test time is competitive to that of dedicated networks trained with those features only. Compared to existing methods such as Mod-Drop and AllFeatNet, the range of performance drops in comparison to dedicated networks is the least in our case. More precisely, the widest drop range for us is up to -2.8% w.r.t. that of LBPNet in AllFeatNetInit network. While for the same feature, it is up to -4.7% in Mod-Drop and up to -21.8% in AllFeatNet w.r.t. that of LBPNet. These results clearly demonstrate that our method is more robust in retaining the performances of individual features while projecting them in common subspace.

Performance comparison with Multi-feature Networks.

Table 3 compares the mean average precision (mAP) of different multiple features based networks with the proposed method. For a network with 3 different types of input features, there are 7 different possible combinations of feature(s) at test time. The performance shown in the table is the mean AP obtained with all these combinations. The proposed method outperforms the other multi-feature-

Features	Dedicated Network	AllFeatNet	Mod-Drop	Ours (FVNetInit)	Ours (AllFeatNetInit)
FV	69.0%	64.2% (-4.7%)	70.0% (+1%)	68.7% (-0.3%)	68.8% (-0.2%)
CNN	68.7%	63.3% (-5.5%)	68.2% (-0.5%)	68.1% (-0.6%)	67.9% (-0.8%)
LBP	64.3%	42.5% (-21.8%)	59.6% (-4.7%)	62.1% (-2.2%)	61.5% (-2.8%)

Table 4: Comparing the proposed methods with other methods using dedicated networks. The table shows that the performance of the proposed methods is competitive to the one of dedicated networks, while the performance of other compared methods is significantly low, particularly in the case of LBPs.

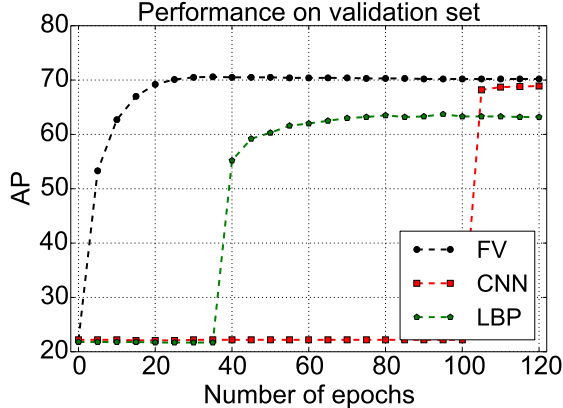


Figure 4: Performance of FVs, CNNs, and LBPs measured on the validation set.

based networks. This shows that the proposed network and the multi-stage training strategy is capable of making better predictions in the presence of more information i.e. multiple types of features at a time and are optimal to every combination of features.

Fig. 3 shows the performance comparison between the proposed methods with AllFeatNet at different levels of feature combinations. From the bar-chart, we can observe that, when all the features are available at test time, AllFeatNet performs better than ours. It is expected too, because this approach is optimized only for this combination. But this is the most unlikely scenario for the applications we are addressing, due to constraints such as computing resources and time, etc. Out of other 6 cases, our method performs substantially better and gives similar performance in one case. This shows that our method leverages all the features available and when more information is present, gives better performance. Unlike AllFeatNet, the proposed method is optimal in every combination of features too.

4.5 Qualitative results

Fig. 5 shows the qualitative performances comparison between the baselines and the proposed method. We randomly choose three different test images and used them for evaluation. Here, we consider LBPs (the simplest feature type) only for evaluation. Thus for both the single feature network (LBPNet) and multi-feature network (AllFeatNet and ours), only LBPs are available at test time. In the figure we can see the top 7 attributes predicted by the compared methods. For each of the attributes, the corresponding score shows the probability of an attribute being present in the given image. On the basis of the number of correct predicted attributes, the performances of LBPNet and the pro-

posed method is comparable in two cases (first two cases). While in the third case, our method (4 correct predictions) is even better than LBPNet (3 correct predictions). This further validates that the proposed method retains the property of component features. The performance of AllFeatNet is comparatively poorer than LBPNet and ours for all test images. Moreover, it is important to note that the scores corresponding to the predicted attributes by AllFeatNet are small. This suggests that with this approach the predictive power of LBPs is masked by other strong features e.g. FV and CNNs.

5. CONCLUSIONS

We propose a novel hybrid deep neural network and a multistage training strategy, for facial attribute classification. We demonstrated, with extensive experiments, that the proposed method retains the performance of each of the component features while aligning and merging all the features in the same subspace. In addition to it, when more than one feature type is present, it improves the performance and attains state-of-art performance. The proposed method is also easily adaptable to new features simply learning the feature specific parameters. This avoids retraining the existing network. Since the majority part of the network is shared among all the feature types, the proposed method reduces the number of parameters.

Acknowledgments This project is funded in part by the ANR (grant ANR-12-SECU-0005).

6. REFERENCES

- [1] Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)
- [2] Perronnin, F., Larlus, D.: Fisher vectors meet neural networks: A hybrid classification architecture. In: CVPR. (2015)
- [3] Bhattarai, B., Sharma, G., Jurie, F., Pérez, P.: Some faces are more equal than others: Hierarchical organization for accurate and efficient large-scale identity-based face retrieval. In: ECCV Workshops. (2014)
- [4] Bhattarai, B., Sharma, G., Jurie, F.: CP-mtML: Coupled projection multi-task metric learning for large scale face retrieval. In: CVPR. (2016)
- [5] Sharma, G., Jurie, F.: Local higher-order statistics (LHS) describing images with statistics of local non-binarized pixel patterns. CVIU (2016)
- [6] Simonyan, K., Parkhi, O.M., Vedaldi, A., Zisserman, A.: Fisher vector faces in the wild. In: BMVC. (2013)
- [7] Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition. In: BMVC. (2015)



LBPNet			AllFeatNet		Ours(AllFeatNetInit)	
	blond hair	0.43	mouth stly open	0.17	wavy hair	0.88
	pointy nose	0.43	black hair	0.20	attractive	0.93
	Attractive	0.56	oval face	0.26	bushy eyebrows	0.94
	heavy makeup	0.70	pointy nose	0.39	heavy makeup	0.96
	w. lipstick	0.94	young	0.72	young	0.99
	young	0.95	no beard	0.99	w. lipstick	0.99
	no beard	1.00	male	0.99	no beard	1.00
	straight hair	0.38	blurry	0.16	attractive	0.85
	attractive	0.57	w. necktie	0.17	male	0.92
	black hair	0.61	mouth stly open	0.28	bushy eyebrows	0.94
	male	0.69	black hair	0.76	no beard	0.98
	young	0.88	young	0.81	black hair	0.99
	bangs	0.98	no beard	0.99	bangs	0.99
	no beard	0.99	male	1.00	young	0.99
	bags under eyes	0.23	high cheekbones	0.11	oval face	0.55
	oval face	0.28	pointy nose	0.27	mouth stly open	0.72
	male	0.53	oval face	0.41	bushy eyebrows	0.80
	young	0.69	young	0.56	no beard	0.85
	mouth stly open	0.70	mouth stly open	0.86	young	0.88
	w. hat	0.96	no beard	0.96	w. hat	0.93
	no beard	0.98	male	1.00	male	0.98

Figure 5: Qualitative results comparison of the proposed method with other methods. Top 7 attributes predicted by these methods are shown. As before green color indicates relevant attributes whereas red color indicates irrelevant attributes for the image. (Better viewed in color)

- [8] Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR. (2014)
- [9] Bhattarai, B., Sharma, G., Lechervy, A., Jurie, F.: A joint learning approach for cross domain age estimation. In: ICASSP. (2016)
- [10] Guo, G., Zhang, C.: A study on cross-population age estimation. In: CVPR. (2014)
- [11] Learned-Miller, E., Huang, G.B., RoyChowdhury, A., Li, H., Hua, G.: Labeled faces in the wild: A survey. In: Advances in FDFIA. (2016)
- [12] Uricchio, T., Bertini, M., Seidenari, L., Bimbo, A.: Fisher encoded convolutional bag-of-windows for efficient image retrieval and social image tagging. In: ICCV Workshops. (2015)
- [13] Kumar, N., Belhumeur, P., Nayar, S.: Facetracer: A search engine for large collections of images with faces. In: ECCV. (2008)
- [14] Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV. (2009)
- [15] Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV. (2015)
- [16] Zhang, N., Paluri, M., Ranzato, M., Darrell, T., Bourdev, L.: Panda: Pose aligned networks for deep attribute modeling. In: CVPR. (2014)
- [17] Kang, S., Lee, D., Yoo, C.D.: Face attribute classification using attribute-aware correlation map and gated convolutional neural networks. In: ICIP. (2015)
- [18] Neverova, N., Wolf, C., Taylor, G., Nebout, F.: Moddrop: adaptive multi-modal gesture recognition. In: PAMI. (2015)
- [19] Kahou, S.E., Pal, C., Bouthillier, X., Froumenty, P., Gülccehre, cC., Memisevic, R., Vincent, P., Courville, A., Bengio, Y., Ferrari, R.C., et al.: Combining modality specific deep neural networks for emotion recognition in video. In: ICMI, ACM (2013)
- [20] Srivastava, N., Salakhutdinov, R.R.: Multimodal learning with deep boltzmann machines. In: NIPS. (2012)
- [21] Wu, Z., Jiang, Y.G., Wang, J., Pu, J., Xue, X.: Exploring inter-feature and inter-class relationships with deep neural networks for video classification. In: ICM, ACM (2014)
- [22] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML. (2011)
- [23] Vedaldi, A., Fulkerson, B.: VLFeat: An open and portable library of computer vision algorithms. <http://www.vlfeat.org/> (2008)