EMPIRICAL ANALYSIS OF CASCADE DEFORMABLE MODELS FOR MULTI-VIEW FACE DETECTION

J. Orozco¹, B. Martinez¹ and M. Pantic^{1,2}

¹Department of Computing, Imperial College London, UK

²Faculty of Electrical Engineering, Mathematics and Computer Science,
University of Twente, The Netherlands

ABSTRACT

In this paper, we present a face detector based on Cascade Deformable Part Models (CDPM) [1]. Our model is learnt from partially labelled images using Latent Support Vector Machines (LSVM). Recently Zhu et al. [2] proposed a Tree Structure Model for multi-view face detection trained with facial landmark labels, which resulted on a complex and suboptimal system for face detection. Instead, we adopt CDPMs enhanced with a data-mining procedure to enrich models during the LSVM training. Furthermore, a post-optimization procedure is derived to improve the performance of the CDPMs. Experimental results show that the proposed model can deal with highly expressive and partially occluded faces while outperforming the state-of-the-art face detectors by a large margin on challenging benchmarks such as the Fddb [3] and the AFLW [4] databases.

Index Terms— Multi-View Face Detection, Cascade Deformable Models, Fddb database, AFLW database.

1. INTRODUCTION

Multi-View Face Detection (MVFD) has been a challenging topic over the last decade [5]. The Viola&Jones (V-J) [6] face detector has been a milestone in the state-of-the-art making face detection feasible in real world applications. It provides reliable performance for head pose rotations up to 30° of yaw and 15° of pitch. They later proposed a two-stage MVFD, which first estimates the face pose and then evaluates the face detector of the estimated pose [7].

Several works have been proposed based on the V-J framework. In [8], a face detector using FloatBoost features, a floating search AdaBoost and a pyramid structure was presented. The method can deal with non-frontal faces, a smaller set of features is required and it is faster than V-J. However, this commercial system requires five times more training time than V-J. Real AdaBoost was used in [9] to

train a view-based classifier using Haar-likes. This work was extended in [10], where a Vector Boosted algorithm and a pyramid cascade were combined to outperform V-J. A large performance improvement was obtained by substituting the Haar-likes for SURF features in [11]. This work has shown the best performance to date over the Fddb benchmark database [3]. SURF features were used as descriptors to learn weak classifiers using logistic regression. Face detection is then performed by applying a cascade of SURF weak classifiers trained with billions of samples within one hour.

Recently Zhu et al. [2] proposed a Deformable Parts Model (DPM) for joint face detection, pose estimation and facial landmark detection. It is a Tree Structure Model (TSM) composed by 13 head poses and up to 68 part filters per pose corresponding to the facial landmarks. This approach showed a better performance than V-J methodologies for MVFD in constrained conditions. This was due to a finer face representation based on HOG features and view-dependent models leading to a better discrimination. However, this work was aimed for facial landmark detection while being suboptimal when only face detection is intended. Firstly, it requires an exhaustive facial landmark labelling, which reduces the amount of training data that can be used. Secondly, learning and searching a tree-structure make the algorithm too slow for face detection in practical applications. Finally, it is limited to high resolution images as the part filters rely on local statistics for a successful detection.

We argue that the baseline framework on DPMs as defined in [12] is more suitable for MVFD than the TSMs [2], which is derived from [12]. These star-structured models have shown strong detection performance on difficult benchmarks such as the PASCAL datasets [13]. Star models can represent rigid and non-rigid facial textures by using mixtures of multi-scale DPMs. Outperforming models are obtained by combining Latent Support Vector Machines (LSVM) and data-mining procedures. Finally, a Cascade Deformable Part Model (CDPM) [1] can speed up over 20 times the DPM's detection without sacrificing detection accuracy.

In this paper, we present an empirical analysis of CDPMs to address the problem of reliable MVFD. First, we describe

Acknowledgments This work has been supported by the European Research Council under the ERC starting grant agreement no. ERC-2007-StG-203143 (MAHNOB) and the European Community's 7th Framework Programme [FP7/20072013] under grant agreement no. 231287 (SSPNet).

a data-mining process to incrementally learn DPMs from partially labelled data using the LSVM algorithm. Second, we derive a post-optimization procedure for the CDPMs training that improves significantly its performance. As a result, we obtain a face detector that outperforms the state-of-the-art face detectors on challenging benchmark datasets such as the FDDB [3] and the AFLW [4].

2. CASCADE DEFORMABLE PART MODELS

2.1. Deformable Part-Based Models

A DPM with n parts is defined as $\beta = (r, c_1, \dots, c_n, b)$, where r is a coarse-scale global root filter, c_i is a model for the i^{th} part and b is a bias term. Part filters are defined as (f_i, v_i, d_i) , where f_i is a fine-scale at twice the resolution of the root filter. The spatial distribution of part filters is defined by both v_i and d_i , the anchor and deformation penalty, respectively. Fig.1 shows an example of a 4-Pose DPM with six part filters per component.

Fig. 1. Right view components of a 4-Pose DPM. First two images are the frontal components, root and six part filters, while the last two images are the profile components.

Our model is trained from partially labelled data using a LSVM algorithm [12]. Face images are labelled with bounding boxes, which are used to build the feature model of the root filter. More complete labelling might be used such as the facial landmarks used by Zhu et al. [2], which resulted in a complex model for face detection due to the use of suboptimal parts. Instead, we treat part locations as latent variables during training, i.e. they are automatically detected using the root filter. To illustrate this, let us consider a model β scoring an example x with a function of the form:

$$S_\beta(x) = \Phi(r) + \sum_{i=1}^n \max_{\delta_i \in \Delta} \Phi(c_i, \delta_i) - d_i(\delta_i) \quad (1)$$

where $\Phi(r)$ is the root filter response, δ_i gives the displacement of part filters relative to its anchor and root's position. Thus, $\Phi(c_i, \delta_i) - d_i(\delta_i)$ scores the contribution of the part filters over displacements and the deformation cost associated with the displacement.

The LSVM is discriminatively trained using an objective function $L_D(\beta) = 0.5 * ||\beta||^2 + C \sum_{i=1}^k \max(0, 1 - y_i S_\beta(x_i))$, where $D = \{(x_1, y_1), \dots, (x_n, y_n)\}$ is the training set, C is a regularization term and $y_i \in \{-1, 1\}$ are the binary labels.

To train a face detector with high performance, LSVM relies on the root filters to learn the part filters as latent variables. To this end, we propose to split the positive training set, D_p , into easy and hard positives, D_{ep} and D_{hp} , respectively. Thus, LSVM learns a coarse root filter using D_{ep} , which is used then to re-score the examples in D_{ep} and obtain a set of latent values, Z_{ep} . Next, LSVM enriches the root filter by minimizing the objective function $L_{D_{ep}}(r, Z_{ep})$ using both labelled and latent variables. Finally, a similar data-mining process is used with the negative examples to ensure the root filter has high precision-recall.

Part filters are learnt using all the positive examples, D_p , the corresponding latent variables, $Z_p = \{Z_{ep} \cup Z_{hp}\}$, and the objective function $L_{D_p}(\beta, Z_p)$. Latent part-locations, $z_p \in Z_p$, are computed at twice the resolution of the root filter and scored by the Eq.1. Thereby, part filters are built using higher resolution features computed over highly scored latent positive examples. Consequently, the root filter captures coarse resolution edges such as the face boundary while the part filters capture details such as eyes, nose and mouth.

2.2. CDPM Training

Given a DPM trained a Star-Cascade (SC) algorithm [1] may be applied to speed-up the detection without loss of accuracy. To this end, a CDPM is trained to find hypothetical object locations that are later validated by the DPM. Although this procedure is not specific of the star model, the CDPM root filter is tailored to scan the image at a low resolution whereas part filters are used at high resolution over the locations provided by CDPM's root filter.

The SC algorithm learns a global threshold, T , to score the most likely locations with the CDPM's root filter, $S_c(r) \geq T$. These scores are accumulated throughout the cascade's stages upon the contribution of the part filters to the detection. If $S_c(r)$ with the first i parts is lower than a threshold τ_i , the root location is not evaluated for the rest of the cascade, this is known as *hypothesis-pruning*. SC will also skip locations if the deformation d_i is below a threshold τ'_i . Finally, the SC algorithm will use the CDPM for hypothesis-pruning at early stages, but the DPM is used at later stages to re-scan the image at the most likely locations.

Aiming to speed-up the detection, simplified CDPMs were obtained in [1] by using PCA-HOG features for the root filter. These are HOG features projected onto the first 5 eigenvectors. Here, we propose a post-optimization procedure to improve the CDPM's performance: this is, to compute the 5-PCA-HOG features from both labelled and latent (easy and hard) positive examples, D_p and Z_p .

3. EXPERIMENTS

We trained a 4-Pose CDPM using images for near-frontal, $[0^\circ, 30^\circ]$, and profile $(30^\circ, 90^\circ]$ faces. Face images were

Fig. 2. Performance on the Fddb. Top and bottom are the discrete and the continuous ROC curves, respectively. The TPR is reported for FP=200 along with each method.

flipped to build symmetric models. We used 35,738 face images labelled with bounding boxes, see Table 1. Specifically, we first learnt a 4-Pose DPM using the LSVM algorithm as explained in Section 2.1. Expressive images were used as easy positives, D_{ep} , whereas AFLW images were used as hard positives, D_{hp} . Images were clustered into near-frontal and profile using the 3D head pose estimation given by the tracking system in [14]. The training set only contained faces with pitch and roll angles lower than 20° .

database	# Examples
AFLW [4]	10,096
Cohn-Kanade [15]	3,130
DaFeX [16]	996
FGnet [17]	1,962
MMI [18]	1,150
Mind Reading [19]	6,552
MultiPIE [20]	7,952
Head Pose Database [21]	3,900

Table 1. Training datasets with 35,738 positive examples.

3.1. MVFD with Viola&Jones

We trained a VJ-MVFD face detector as the work in [7] is not publicly available. The training has been carried out using the OpenCV library [22], a Gentle AdaBoost classifier, the upright Haar-like features and a tree-based cascade structure for an efficient search [23]. We trained a 6-Pose MVFD is trained for near-frontal ($[0^\circ, 30^\circ]$), half-profile ($[30^\circ, 60^\circ]$) and full profile ($[60^\circ, 90^\circ]$) faces. The training set of 35,738 face images from Table 1 was extended to 100,000 positive examples by flipping the images and applying random distortions. Our training of the VJ-MVFD took up to four weeks per pose with the final configuration.

To detect a face, the VJ-MVFD runs all pose-specific de-

tectors in parallel. Next, detections are merged by first using a disjoint-set data structure function [22] to cluster the detected rectangles according to their size and location. Then, clusters with a minimum number of rectangles are eliminated. Finally, a non-maximum suppression function is used to merge the remaining detections. The detections are scored as the maximum response among the pose-specific detectors.

3.2. Experiments on Fddb

The Fddb database [3] is the latest benchmark dataset for face detection in real world scenarios. It contains 2,845 images and 5,171 faces acquired under unconstrained conditions. We report the performance on the Fddb according to the evaluation scheme proposed by Jain et al.[3]. Fig.2 shows both the discrete and the continuous ROC curves for our two methods, the 4-Pose CDPM-MVFD and the VJ-MVFD. In addition, we compare their performance against the TSM method [2] and the top five face detectors reported on the Fddb [24] including the VJ-OpenCV implementation for frontal faces, see Fig.2.

It can be seen from Fig.2, both discrete and continuous ROC curves, that our 4-Pose CDPM-MVFD achieves the highest performance on the Fddb. The True Positive Rate (TPR) is greater than all methods at any rate of false positives. Specifically, we compare the TPR at a small number of false positives such as 200. At this point, the 4-Pose CDPM-MVFD improves the TSM in more than 60%, 45% is attained with respect to the VJ-MVFD and 13% over the face detector by Li et al.[11], which is the best performance reported to date on the Fddb. Note that the 4-Pose CDPM-MVFD can recall up to 92.96% of the faces in the Fddb. By contrast, the TSM [2] method can at most recall 59.16% of the faces.

On the other hand, the 6-Pose VJ-MVFD can just perform as well as the VJ-OpenCV, which uses only a frontal classifier without filtering neighbouring detections.

Fig. 3. Detection examples using the 4-Pose CDPM-MVFD on FDDB and AFLW databases. Red boxes are used for highly scored detections while the green boxes are for low scores. The blue boxes within the red boxes are the part filters locations.

3.3. Experiments on AFLW

The AFLW database contains 24,686 faces in 21,328 images, with manually annotated facial landmarks. The 3D face pose can be estimated by fitting a 3D face model to the provided landmarks. The database is released in three folders, such that testing images are taken from the first two folders and training images from the third folder. Fig. 3 shows detection examples of out MVFD on the FDDB and AFLW databases.

can recall 65.12% and 78.01%, respectively.

The AFLW database contains faces with larger head poses and higher resolution than the FDDB. Therefore, the VJ-MVFD fails on profile faces whereas the TSM can deal with both pose variation and high resolution. The FDDB is the most challenging benchmark due to low resolution and occluded faces. Consequently, the TSM performed better on the AFLW whilst VJ-MVFD performed better on the FDDB.

Fig. 4. Performance on the AFLW. Discrete ROC curves are compared according to the TPR with at most 10% of FPR.

The AFLW testing set contains 14,675 images and 17,166 annotated faces. In this experiment, we compare the face detection results of the 4-Pose CDPM-MVFD with the VJ-MVFD and the TSM face detector. Like in the previous section, we compare the face detection performance according to the discrete ROC curves at a maximum False Positive Rate (FPR) of 10%, see Fig.4. Again, our 4-Pose CDPM-MVFD outperforms both VJ-MVFD and TSM face detectors with margins of 57% and 15%, respectively. Furthermore, the 4-Pose CDPM-MVFD can recall up to 95.08% of the faces on the AFLW, whilst the VJ-MVFD and TSM

3.4. Detection Speed

Although we are not aiming a real time face detector, we have obtained a MVFD that is comparable in speed to the VJ-MVFD. We tested both the 4-Pose CDPM-MVFD, the VJ-MVFD and the TSM scanning the 2,845 FDDB images, which have an average resolution of 377x399 pixels. Our model reported an average detection time of 0.46 seconds, the VJ-MVFD took an average time of 0.52 seconds whilst the TSM achieved an average time of 26.06 seconds. Moreover, this detection speed also contributes to a fast training process when the LSVM does data-mining over both positive and negative examples. The training of both DPM and CDPM may take between 24 and 48 hours.

4. CONCLUSIONS

This paper presents an empirical analysis of two methods for MVFD on unconstrained and challenging databases. The experiments show that the CDPMs method [1] can be applied to learn an efficient MVFD. We enrich the model by discriminatively training the LSVM with easy and hard-positive data. Furthermore, we trained CDPMs with both labelled and latent (easy and hard) positives to improve their performance. Experimental results show that our face detector significantly outperforms the state-of-the-art methods by a large margin. Lastly, we provide code for Matlab and C++ for reproducing our experiments. It can be found at <http://ibug.doc.ic.ac.uk/resources>

5. REFERENCES

- [1] P. Felzenszwalb, R. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *CVPR*. IEEE, 2010, pp. 2241–2248.
- [2] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*. IEEE, 2012, pp. 2879–2886.
- [3] V. Jain and E. Learned-Miller, "FDDB: A benchmark for face detection in unconstrained settings," *University of Massachusetts, Amherst*, 2010.
- [4] M. Koestinger, P. Wohlhart, P.M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *BeFIT*, 2011.
- [5] C. Zhang and Z. Zhang, "A survey of recent advances in face detection," *Microsoft Research*, June, 2010.
- [6] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *CVPR*. IEEE, 2001, vol. 1, pp. I–511.
- [7] M. Jones and P. Viola, "Fast multi-view face detection," *Mitsubishi Electric Research Lab TR-20003-96*, vol. 3, 2003.
- [8] S. Li, L. Zhu, Z.Q. Zhang, A. Blake, H.J. Zhang, and H. Shum, "Statistical learning of multi-view face detection," in *ECCV*. Springer, 2006, pp. 117–121.
- [9] B. Wu, H. Ai, C. Huang, and S. Lao, "Fast rotation invariant multi-view face detection based on real adaboost," in *FG*. IEEE, 2004, pp. 79–84.
- [10] C. Huang, H. Ai, Y. Li, and S. Lao, "Vector boosting for rotation invariant multi-view face detection," in *ICCV*. IEEE, 2005, vol. 1, pp. 446–453.
- [11] J. Li, T. Wang, and Y. Zhang, "Face detection using surf cascade," in *ICCV*. IEEE, 2011, pp. 2183–2190.
- [12] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *TPAMI*, vol. 32, no. 9, pp. 1627–1645, 2010.
- [13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL VOC2012 Results," .
- [14] F. Dornaika and J. Orozco, "Real time 3d face and facial feature tracking," *Journal of Real-Time Image Processing*, vol. 2, no. 1, pp. 35–44, 2007.
- [15] T. Kanade, J.F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *FG*. IEEE, 2000, pp. 46–53.
- [16] A. Battocchi, F. Pianesi, and D. Goren-Bar, "Dafex: Database of facial expressions," *Intelligent Technologies for Interactive Entertainment*, pp. 303–306, 2005.
- [17] F. Wallhoff, "Fgnet-facial expression and emotion database," *Technische Universität München*, 2004.
- [18] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*. IEEE, 2005, pp. 5–pp.
- [19] W. Juneke, "Mind reading: The interactive guide to emotions," *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, vol. 16, no. 4, pp. 182, 2007.
- [20] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.
- [21] N. Gourier, D. Hall, and J.L. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *ICPR*, 2004, pp. 1–9.
- [22] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.
- [23] R. Lienhart, A. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," *Pattern Recognition*, pp. 297–304, 2003.
- [24] V. Jain, "FDDB Results," <http://vis-www.cs.umass.edu/fddb/results.html>.