Where is my Friend? - Person identification in Social Networks

Deepak Pathak[1,2], Sai Nitish Satyavolu[3] and Vinay P. Namboodiri[2]
[1] Department of Electrical Engineering & Computer Sciences, UC Berkeley
[2] Department of Computer Science and Engineering, IIT Kanpur
[3] Department of Electrical Engineering, IIT Kanpur

***Abstract*— One of the interesting applications of computer vision is to be able to identify or detect persons in real world. This problem has been posed in the context of identifying people in television series [2] or in multi-camera networks [8]. However, a common scenario for this problem is to be able to identify people among images prevalent on social networks. In this paper we present a method that aims to solve this problem in real world conditions where the person can be in any pose, profile and orientation and the face itself is not always clearly visible. Moreover, we show that the problem can be solved with as weak supervision only a label whether the person is present or not, which is usually the case as people are tagged in social networks. This is challenging as there can be ambiguity in association of the right person.**

The problem is solved in this setting using a latent max-margin formulation where the identity of the person is the latent parameter that is classified. This framework builds on other off the shelf computer vision techniques for person detection and face detection and is able to also account for inaccuracies of these components. The idea is to model the complete person in addition to face, that too with weak supervision. We also contribute three real-world datasets that we have created for extensive evaluation of the solution. We show using these datasets that the problem can be effectively solved using the proposed method.

I. Introduction

Given an image with several people, as humans, we are easily able to identify whether a particular person (a friend) is present or not. In this paper, we present a method that is able to achieve the same, i.e. given an image, we are able to identify whether a particular person is present or not. Further we are able to localize and draw a bounding box around the person. This is achieved by using as supervision only a set of images that has the person (the positive set) and a set of images that does not have the person (the negative set). This setting is commonly termed weakly supervised setting as no information is provided about the location of the person or specific information about the person. However, this setting is more practical as tags are widely prevalent in common social networking settings, however, labeling specifically being more laborious is less utilized by users of social networks. It is also important to distinguish this problem from the problem of Person Re-identification, where the aim is to identify the same person in multiple images (for eg., from different non-overlapping cameras). The key point to note is that Person Re-identification does not disambiguate the person of interest with the detected person, which our algorithm does.

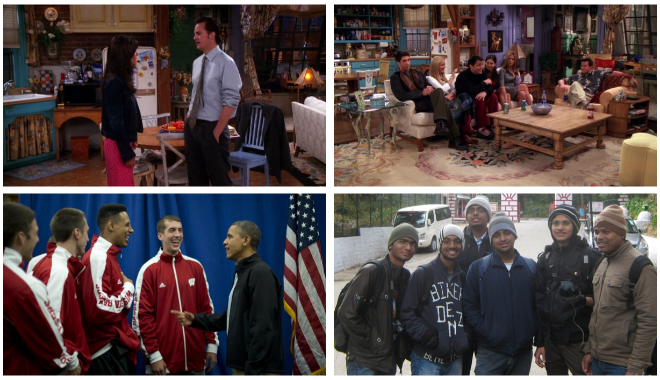


Fig. 1: Sample positive images from the datasets. Images in the top row are from TV Series Dataset, bottom left image is from Obama Dataset and last image is from Social Network Dataset.

Previously, the successful methods for person identification have relied primary on facial features [2] or multiple-view similarity [7], [16]. This has been shown to work for videos where the face is detected using the Viola Jones face detector [18]. However, this approach is not directly applicable for images as face detection for faces that are not frontal is not as successful. Real world images contain people in varied pose, profile and orientation where the face itself is not always completely visible as shown in Figure 1. Further, for video based person identification methods, we can rely on a large number of positive samples. The task addressed in this paper relies on access to limited set of distinct images for supervision as is common for social network images where the person of interest is tagged. This setting has not been explored in previous related work. Moreover, the earlier methods rely on strong supervision where explicit instances of faces of people have to be provided resulting in the method being laborious. In contrast, the proposed method can work with only tags provided for images of people without explicit location information being provided.

The proposed method is implemented by building up on previous work on person detection by Felzenszwalb [5]. This is more robust than frontal face detection (which we also use) in detecting potential candidate bounding boxes for a specific person. However, just person detection would not help us in the task of weakly supervised person identification as there can be ambiguity in association of the right person. We are able to solve the problem by using a latent parametric

This work was performed when Deepak Pathak was partly at IIT Kanpur.

approach where a latent support vector machine is used [19] and the specific person is a latent parameter in the system.

In the next section, we discuss related previous work. In section III we present the proposed method. We then present the dataset used for evaluation in section IV. Results for the experiments are presented in section V following which we conclude the paper.

II. Previous Work

There has been some interest in the task of person identification. This has however been mainly in the context of naming of characters in TV series [2] or identification of people in multi-camera settings [11]. In this paper, we consider the identification of people from a limited number of widely varying images that are weakly labeled with presence or absence of a person. This is in contrast to the previous methods which make use of videos for person identification. This identification task, while being practically very useful for tagging of people in images, is quite challenging.

In previous methods, there has been work by Apostoloff and Zisserman [2] where the authors have proposed a method to use face detection in videos to be used for person identification. This builds up on face detection methods [18] that is made more robust by kernel based regressors for tracking. The face is described in terms of its parts and is classified using a random ferns classifier. The main challenge in this work is the task of face identification. The task of face identification has been addressed by Guillaumin [9] by using metric learning. They propose techniques for metric learning and classification such as logistic discriminant metric learning and marginalized k-nearest neighbors. Very recently the task of person identification has been considered by Beuml *et al.* [3] where they construct multinomial logistic regression classifiers for multi-class face recognition in the semi-supervised setting where not all face-tracks have annotation. This method is then evaluated on two TV series.

However, in contrast to the above mentioned methods where the challenge is to verify the identity of a given face, we have to be able to identify the person. As we are considering the task of person identification in images, we aim to do so even in cases where the face is not fully visible for a diverse range of person appearances. There has been some work aimed towards the task of person identification in video [17]. In the work by Tapaswi *et al* [17], the authors consider a probabilistic framework where they propose the use of a Markov random field to model each TV series that integrates face-recognition, clothing appearance, speaker recognition and contextual constraints. However, as previously mentioned this method is proposed for video, specifically TV series and is not directly applicable in our setting.

There has also been work that addresses the task of person re-identification in multi-camera networks [11]. In recent work by Liu *et al*, the authors consider this task and specifically propose a method that is able to obtain a ranking of persons by observing people in a multi-camera network setting. In [13], the authors propose a novel SVM-based ranking method for person re-identification. Zheng *et al* [20] propose a method for learning the optimal similarity measure between a pair of images by formulating the person re-identification problem as a relative distance comparison (RDC) learning problem. Person re-identification only gives the correspondences in multiple images [6]. Unlike our methodology, it does not disambiguate the identity of the detected person if the identity of the person in the gallery image is unknown. While, there is variation in the appearances of people in multiple cameras, as most such methods consider the case where the person is being observed between multiple cameras, there are significant commonalities such as clothing being consistent that does aid solving this task. This commonality is not present in our task, which makes the problem harder.

As the method for person identification from a small number of images has not directly been addressed, we propose our method that does so and additionally we also present three real-world datasets drawn from TV series, celebrity and social-network for thoroughly evaluating the proposed method. From the evaluation, we observe that the method that integrates state-of-the-art feature description with a principled framework does manage to significantly solve both the classification and localization tasks. In the next section, we present our method.

III. Method

Our aim is to detect and localize a target person of interest (a friend) in a test image. The challenge that is addressed here is to resolve the ambiguity in person identification when we are provided with only information regarding presence or absence of the person in an image. In order to learn a classifier from this dataset, we need to address the issues associated with it. One could run a person detector and obtain bounding boxes corresponding to persons or face-detection and bounding boxes corresponding to faces. However, there would be two kinds of ambiguity present. One kind of ambiguity lies in that of deciding which of the bounding boxes in each image corresponds to the friend. The second kind of ambiguity lies in that in order to ensure robust detection of persons we would be lowering the detection threshold and some of the bounding boxes would indeed not correspond to a person at all. Thus the two ambiguities can be termed the data association ambiguity and the ambiguity arising from classification accuracy. In order to resolve both these ambiguities we rely on a principled framework where we simultaneously resolve the location and classification of the person. This is achieved by formulating the solution of the problem in terms of a principled max-margin framework where the location is treated as a latent variable. The problem can then be solved by using a latent support vector machine [1], [19].

We first formulate the solution of the problem and discuss the learning and inference procedure. The learning and inference are aided by providing sparse set of candidate location hypotheses that are obtained by using object and face detectors [5], [18]. These hypotheses are described by using the current state-of-the-art mid-level feature description using convolutional neural networks that have been trained on image-net dataset [15]. This pipeline of obtaining a sparse set of candidates that are effectively represented enables us to solve the problem. The solution of this problem requires a training set of images with the labels $+1$ or -1. The label $+1$ specifies that the person is present in the image, but not the person's location. The label -1 specifies that the person is absent. Given this data as input the problem is solved irrespective of the number of people present in the training images, the pose of the person, the appearance of the person and other challenges such as occlusion. We now outline the solution of the problem in the following sub-sections.

A. *Learning and Inference*

We are provided with a set (x_i, y_i) consisting of the images x_i and the associated binary label y_i indicating presence or absence of the person. Given a new test image x_j we have to then predict the presence or absence of the person, i.e. predict y_j. This task is a binary classification task. We obtain a representation $\phi(x_i, y_i)$ that jointly represents the relation between input and output. The binary classification task is solved by a discriminant rule of the form

$$f_w(x) = \text{argmax}_y \left[w \cdot \phi(x_i, \hat{y}_i)\right], \tag{1}$$

where w is the parameter vector that is learned from the data. However, the presence or absence of the person depends on an additional parameter that is not observed. This is the latent parameter h_i that specifies the location of the person in the image. Based on this parameter, the representation function then becomes $\phi(x_i, y_i, h_i)$. However, in the practical setting of our problem we do not have access to the ground-truth for the location of the person in the image. We therefore have to learn an inference rule

$$f_w(x) = \text{argmax}_{y,h} \left[w \cdot \phi(x_i, \hat{y}_i, \hat{h}_i)\right], \tag{2}$$

that simultaneously predicts the value of y and h that maximizes the function $f_w(x)$. Learning this inference rule allows us to simultaneously learn the location and classification jointly. This is achieved by using the formulation of Yu and Joachims [19]. Some clarifications are in order now. The function $\phi(x_i, y_i, h_i)$ is the feature vector obtained by using the convolutional neural network. The candidate hypotheses locations $\hat{h}_i$ that are considered for the inference are obtained by using the person detector. Thus if any person is likely to be present at a location $\hat{h}_i$, then we will consider it as a valid hypotheses to predict the presence or absence of a *specific* person that is under consideration. We presently consider only a linear prediction rule, however, the framework is not restricted to linear relation, but can be extended to consider non-linear kernels using the kernel-trick representation of support vector machines.

The learning of the parameter vector w is done by using the loss-augmented inference method of Yu and Joachims. This is done by solving the following optimization problem

$$\text{min}_w \frac{1}{2}||w||^2 + C\sum_{i=1}^{n}\Big[\text{max}_{y_i,h_i}\Big[w \cdot \psi(x_i, \hat{y}_i, \hat{h}_i) + \triangle(y_i, \hat{y}_i, \hat{h}_i)\Big] - \max_{\hat{h}_i}\Big[w \cdot \psi(x_i, y_i, \hat{h}_i)\Big]\Big] \tag{3}$$

This optimization equation involves multiple terms. In this optimization we measure the difference between the best estimated pair $(\hat{y}_i, \hat{h}_i)$ with respect to the weight vector w and the best estimated $\hat{h}_i$ when the ground truth label y_i is available. This is done with respect to the loss $\triangle(y_i, \hat{y}_i, \hat{h}_i)$ which is basically chosen to maximize area under the curve (AUC) as suggested in [4]. This optimization is solved by alternating between fixing the latent variables and learning the weight vector w and by fixing the weight parameter vector and estimating the best latent variables. This procedure for solving the optimization is called the Concave - Convex procedure (CCCP).

B. *Hypothesis Generation*

In the principled max-margin formulation, we estimate the location information as latent parameter in the model. This latent model is inspired by the belief that including subject specific information and rejecting other insignificant content should enhance the classification accuracy by providing better discriminative score [4]. The latent parameter h is completely specified by the opposite coordinates of bounding box. Although one could optimize the latent parameter over all possible locations, we assist the inference procedure by seeding in the sparse set of locations which have high likelihood of person being present. This is beneficial to both convergence rate and computational requirements of algorithm.

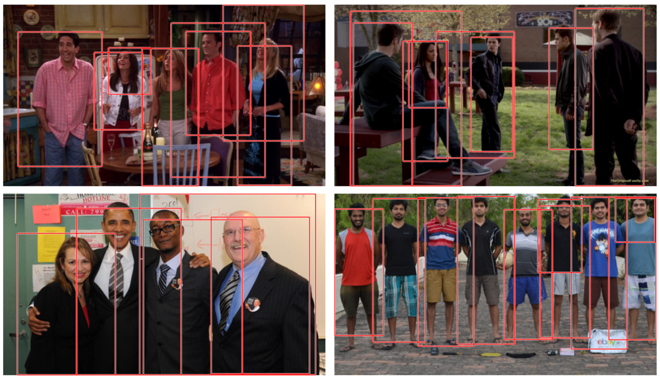
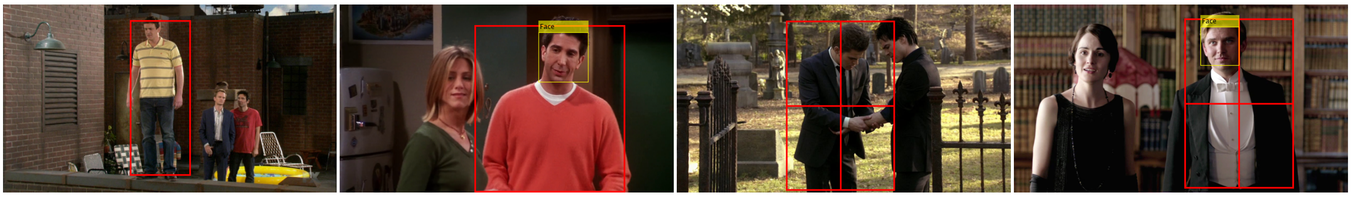


Fig. 2: Sample images from dataset depicting DPM person detection output. Lower threshold results into larger number of detections.

To generate the candidates for location hypothesis $\hat{h}$, we use the Deformable Part Model (DPM) based object detection as proposed in [5]. The DPM contains a 'root filter' representing the object as a whole, a set of 'part filters' each representing a particular part and 'deformation



(a) 1-Level Pyramid (b) 1-Level Pyramid with Face (c) 2-Level Pyramid (d) 2-Level Pyramid with Face

Fig. 3: Illustratve figures for latent models of hypothesis description.

models' for each part. Each part model contains the descriptor information of that particular part, and the deformation cost for each possible placement of that part relative to the root location. The score for a particular configuration is the result of individual filter responses minus the deformation costs. This configuration score is optimized with respect to positioning of mutual parts to obtain the overall score for that root location. The object is then detected by non-maximum suppression and subsequent thresholding of the overall score.

We use the pre-trained root and part models trained on PASCAL VOC 2010 person dataset. For experimental purposes, instead of generating hypothesis using their threshold, we use a lower threshold to conservatively include all possible person detections. As shown in Figure 2, our changed threshold modifies the non-maximal suppression and we end up with lot more hypotheses.

C. Hypothesis Description

For a given hypothesis $\hat{h}$, the feature vector is defined by the descriptor evaluated on the hypothesised location. Recently, convolutional neural networks (CNN) have been increasingly used for feature extraction in various computer vision problems. These biologically inspired models are partially invariant to translation, making them highly suitable for problems like object recognition. For describing the hypothesis, we use OverFeat [15] features, a pre-trained CNN trained on the ImageNet 2012 dataset. The Overfeat network takes as input $3 \times 221 \times 221$ sized images and has 25 layers in total. Prior to extracting the features, each input image is resized to 221×221. The responses of the 24^{th} layer are used for the feature representation, giving us a 4096-length feature vector for each input image. We used the 'accurate' version of network provided on the website[1]. The proposed methodology is not specific to any descriptor. We experimented with dense SIFT features [12], but the mid-level CNN based features seem to outperform them. Moreover, Razavian [14] propose empirical analysis to state that generic descirptors extracted from CNNs are robust to various vision applications including fine grained recognition.

Spatial pyramidal division of images has been shown to significantly increase the performance in object recognition [10]. Thus, we also experiment using a 2-level spatial pyramid representation, in which case, we extract Overfeat features for each patch at each level individually and concatenate them. These are depicted in Figures 3a and 3c.

Person identification techniques have previously relied on just features based on facial information [2]. So, next obvious step is to model face explicitly along with body as a whole. We thus augment the pyramidal hypothesis description with facial features as shown in Figures 3b and 3d. After obtaining the bounding boxes around the humans detected in the image using DPM, we then detect the face inside them. This is achieved using the Viola-Jones face detector [18].

The Viola-Jones face detector classifies images based on Haar-like rectangular features. By using the integral image representation, we can compute each of these features at any scale or location in constant time. However, the exhaustive set of rectangular features is very large - for a 24x24 detection window, there are over 180,000 features. Therefore, the detector selects a very small set of these features using a variant of the AdaBoost algorithm. The classification is done using a cascade of boosted classifiers. Smaller, more efficient classifiers are used to reject most of the sub-windows and the remaining sub-windows are passed on to more complex classifiers. Stages in the cascade are trained using AdaBoost.

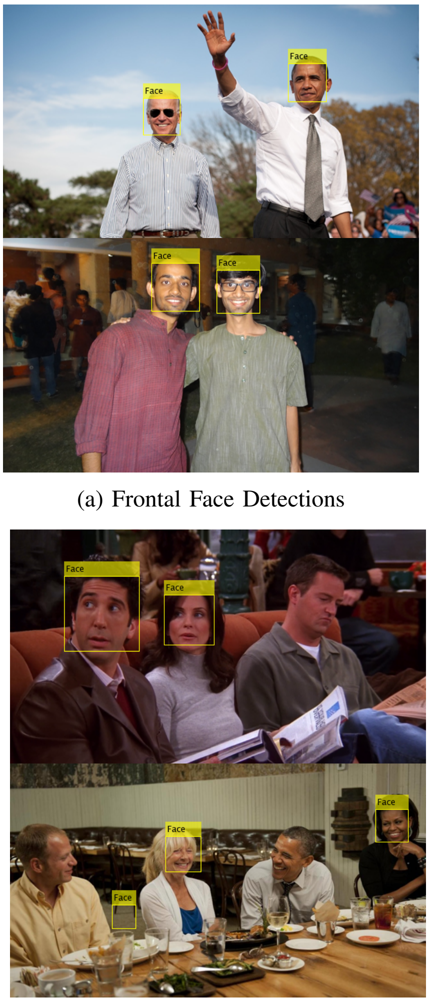
Our aim is to detect the face for each human detected by the DPM. The Viola-Jones face detector, however, poses a problem for our purpose. The detector may not detect the face, or there may be multiple detections for a single human. Table I shows the quantitative result showing the fraction of faces missed by vanilla face detector algorithm in real world datasets. Figure 4 depicts some of such instances from the datasets. To solve this problem, we can use part location information from the DPM. When there are no faces detected, we can take a square patch of a pre-defined size centered around the head location (obtained from the DPM) and consider it to be the face. When there are multiple faces detected, we can choose the face that is closest to the head. However, in our experiments, when there are no detected faces, we take the patch that lies in the center of the top quarter of the image. For the case of multiple detections, we choose the face that lies farthest to the top. We have found that this heuristic works reasonably well in most cases.

D. Algorithm

We now summarize the steps followed by the proposed approach for training.

- For each training image, compute the seed person detection hypotheses using the Deformable Parts

[1] http://cilvr.nyu.edu/doku.php?id=software:overfeat:start



(a) Frontal Face Detections

(b) Missed Face Detections

Fig. 4: Representative figures for the results of Voila Jones Detector.

Dataset	Fraction of faces missed
Chandler Dataset	36.6%
Obama Dataset	39.2%
Social Network Dataset	38.7%

TABLE I: Percentage of human faces not detected by Viola-Jones face detector in the datasets.

Model [5]. Note that this could also be replaced with some other more accurate person detection method.

- Optionally, for each person hypothesis obtain the face detection using the Viola-Jones face detector [18]. As in person detection, we could replace this face detector with some other improved face detection algorithm.
- Obtain for each person detection hypothesis a pyramidal representation using one-level or two level pyramid.
- Describe each cell of the pyramidal representation and the detected face window using overfeat features [15].
- Train and learn the parameter vector w using the latent SVM as described in sub-section III-A. This considers all the features to be represented in the feature vector ϕ where x_i represents the image i, the label y_i represents the presence or absence of the person and the latent variable h_i denotes the various person detection hypotheses.

Once the model parameter w is trained, we can use this at test time to predict for each image the presence or absence of the person. For a test image we similarly compute the person detection hypotheses and describe it using a pyramidal representation and overfeat features. We then compute the prediction using the inference rule given in equation 2. This predicts using each of the person detection hypothesis whether the person is present or absent by computing the maximum score for presence and absence of the person. The score that is maximum is used to predict whether the person is present or absent.

IV. Dataset Description

We have tested our method on three different datasets (see Figure 1). All these datasets contain real world images where the person of interest can be present in any pose, profile and position in the image i.e. non-centered. The first dataset, known as *Chandler dataset* is based on the popular TV show “Friends”. The positive examples are composed of 100 images containing “Chandler”, a popular character from the show. The negative set has a 50% contextual version and a non-contextual version. The 50% contextual version contains 50 contextual images (again from “Friends” but not containing Chandler) and 50 more images from various other TV shows. The non-contextual version contains all the images from other TV shows different from “Friends”. This is to evaluate the role of context in disambiguating the correct person. To furhter validate the efficacy of approach outside the ambience of surroundings present in images taken from “Friends” TV series, we formulate another dataset named ‘Chandler in other movies’. In this, we collect 36 images of the actor who played the role of ‘Chandler’ from different movies.

The second dataset, known as Obama dataset is based on US President Barack Obama. The positive set contains 100 images of Obama and 100 image of other politicians taken from Flickr [2] under Creative Commons license. These images contains scenarios where Barack Obama is present in different profiles and poses etc.

The third dataset, known as Social Network dataset, contains images downloaded from Facebook. The positive set is composed of 100 tagged images of a particular person. The negative dataset of 100 images is also downloaded from Facebook from similar domain but in the absence of person of interest.

The three datasets capture the wide variety of pose, variation, scale, clutter of people in natural image settings.

[2] https://www.flickr.com/photos/barackobamadotcom/

Dataset	1-Level Pyramid	2-Level Pyramid
Chandler Dataset (non-contextual -ve)	84.62 (79.14)	86.61 (81.39)
Chandler Dataset (50% contextual -ve)	68.83 (70.04)	73.01 (75.10)
Obama Dataset	71.25 (74.24)	78.07 (80.14)
Social Network Dataset	86.36 (80.31)	89.67 (82.06)

TABLE II: Baseline Results for *Friend* Classification. The reported values are AUC of precision-recall curve with average precision in parenthesis.

Further, for the social network setting we would be having access to only a limited set of images for training. This represents a significant challenge for person re-identification task. To the best of our knowledge, this represents the first such work addressing this real world task. We have thoroughly evaluated the method on different variations of the task with different settings. While, for the 'Chandler' dataset we have to be able to de-correlate the person from a number of re-occurring people and wide pose variation and scale, for the social network dataset the number of people reoccurring is less, however, each image has a large number of persons of similar appearance which presents the challenge. For, the 'Obama' dataset, the person of interest is visible, but again there are a large number of other persons and there is a wide variety of pose of the person. Note that using face detection alone, we would not be able to identify the person in all these cases. In the next section, we present thorough evaluation of the proposed method for all these datasets. These datasets have been publicly released on the author's website.

V. Experiments and Results

In all our experiments, training is performed on 80% of the data, while testing is done on the remaining 20%. For experimentation purpose, we consider 5 folds of data, and report mean out-of-sample accuracy on test sets corresponding to each fold. The LSVM regularization parameter is chosen to be between 0.5 to 5 in all our experiments.

Classification accuracy, i.e identifying whether the subject (*Friend*) is present in the image or not, is provided in the Table III. For the classification baseline, we extract OverFeat features on the images and classify them using an SVM with Radial Basis Function kernel. The baseline results for classification are shown in Table II. Since we obtained scores for individual examples from max-margin classification, the best method to evaluate performance would be to report area under precision-recall curve (AUC) and average precision (AP). Results are reported for all three datasets with different models of feature extraction.

Accuracy for localization of *Friend* in the images is depicted in similar format in Table IV. Reported accuracy is in 0-1 error format, and is obtained by manual inspection of the resulting output images.

Some of the images have been shown in Figure 5 depicting the results of detection and localization of the person of interest in different datasets. The max-margin classification score is printed on the top-left corner with bounding box around the person of interest.

A. Discussion

As is evident from the results, best performance is achieved using a 2-level pyramidal representation with or without facial features. However, it is important to note that this feature representation is computationally expensive. The next best candidates are the 1-level pyramid with facial features. The 1-level pyramidal representation alone without the facial features gave the least performance, but is also the least computationally expensive of all. It is interesting to see that the main power comes from the facial deep features which should be the case intuitively. The colors of clothes of the person of interest may vary across the dataset with face being the most discriminative part. But it is essential to point out that, in real world setting, person can be present in any pose or profile, thus faces may not be frontal or clearly visible making it difficult for usual face recognition techniques to generalize. This top-down approach, with appropriate heuristics, makes the person identification robust to such instances.

We have compared the proposed method with the one-level pyramid corresponding to the use of Overfeat features [15] over the full image and the two-level pyramid of the overfeat features. This baseline has been shown to be an extremely strong baseline for all visual classification tasks [14]. Compared to the one-level localization setting of the proposed method the one-level pyramid version of the baseline does perform better in the Chandler dataset (contextual and non-contextual). This is because, the ambiguity in person identification is not fully resolved in this setting for this dataset. As the proposed method uses non-convex optimization, we observe that with lesser discriminative features, the wrong association of identity may be possible. The baseline, that uses the full image always contains the person to be identified whereas this may not be the case if the identity which we want to resolve is wrongly associated. However, in the more discriminative two-level feature pyramid, we significantly improve over the baseline. Moreover, the proposed method also solves the localization problem. This is not provided by the baseline.

Recognition accuracy is more for the non-contextual Chandler dataset compared to the 50%-contextual setting, because of difference in ambiance and image quality of the positive and negative images in the former. Intuitively, this means that for the non-contextual setting, the latent-SVM is able to classify the images based on the ambiance and quality, without needing to correctly choose the latent variables. As a consequence, the localization accuracy is less in the non-contextual setting compared to the contextual setting. As an interesting result, this model also scales to the cases when the person playing character of 'Chandler' occurs in different natural settings i.e. other movies.

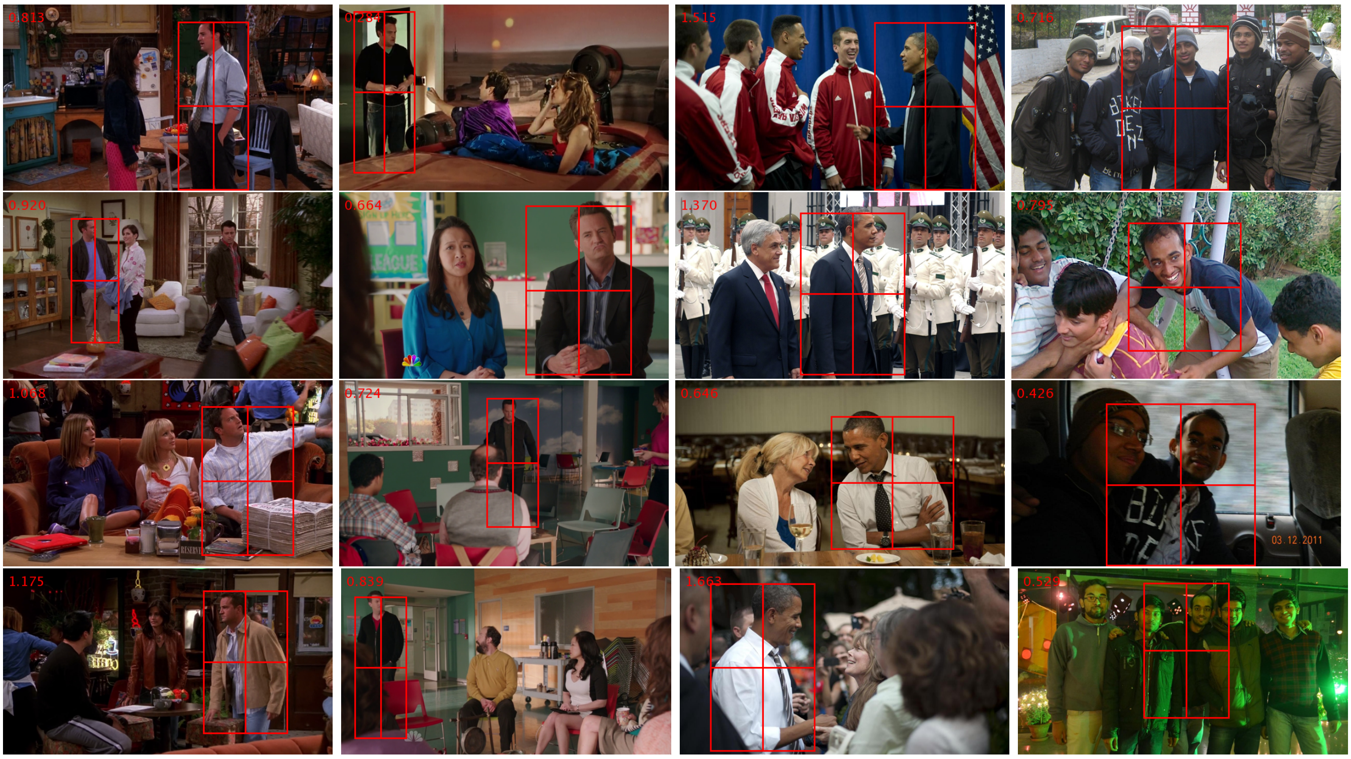
The localization accuracy is lower for the Social Network dataset compared to the other two datasets. Again, this is partly attributed to the positive and negative images being in

Dataset	1-Level Pyramid	1-Level Pyramid with Facial Features	2-Level Pyramid	2-Level Pyramid with Facial Features
Chandler Dataset (v/s non-contextual negative)	85.00 (85.98)	89.20 (88.57)	**90.80** (**92.37**)	89.55 (90.28)
Chandler Dataset (v/s 50% contextual negative)	69.80 (74.43)	**80.40** (**81.74**)	77.10 (79.79)	75.95 (78.96)
Chandler in Other Movies (trained on 'Friends')	57.50 (86.06)	56.11 (73.98)	**68.33** (**80.23**)	63.61 (78.03)
Obama Dataset	72.80 (77.69)	82.40 (85.55)	82.65 (85.49)	**84.75** (**87.71**)
Social Network Dataset	88.35 (89.29)	91.75 (92.76)	**94.85** (**95.17**)	94.55 (94.59)

TABLE III: Results for *Friend* Classification. The reported values are AUC of precision-recall curve with average precision in parenthesis.

Dataset	1-Level Pyramid	1-Level Pyramid with Facial Features	2-Level Pyramid	2-Level Pyramid with Facial Features
Chandler Dataset (v/s non-contextual negative)	56 %	59 %	**76** %	62 %
Chandler Dataset (v/s 50% contextual negative)	54 %	71 %	74 %	**84** %
Chandler in Other Movies (trained on 'Friends')	55.56 %	58.33 %	58.33 %	**63.89** %
Obama Dataset	71 %	**76** %	68 %	72 %
Social Network Dataset	58 %	56 %	58 %	**66** %

TABLE IV: Results for *Friend* Localization. The reported values are percentage of examples where *Friend* is identified correctly.



(a) Chandler Dataset (b) Chandler in Other Movies (c) Obama Dataset (d) Social Network Dataset

Fig. 5: Representative images for the final results of detection and localization in different datasets.

different settings. Also, people in the image tend to stand in close proximity that the DPM sometimes fails to localize all the humans independently.

VI. Conclusion and Future Scope

In this paper, we have addressed the challenging problem in person identification from a limited set of images, as seen in the case of social network settings where images need to be tagged or annotated. While, this problem is of significant interest practically, the commonly prevalent person reidentification methods have surprisingly not yet addressed this challenge. To address the problem of ambiguity in person resolution we have used a principled max-margin framework where the person to be identified is treated as a latent parameter. We make use of the state of the art techniques for person identification and describe the features using discriminative features that are learned using deep convolutional neural networks on imagenet dataset. The evaluation of our method on 3 challenging datasets created by us suggests that while the problem is hard, we can make progress. In future we would like to further improve our method by adding constraints on the solution to ensure higher localization accuracies.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*, pages 561–568, 2002.

[2] N. E. Apostoloff and A. Zisserman. Who are you? – real-time person identification. In *British Machine Vision Conference*, 2007.

[3] M. Beuml, M. Tapaswi, and R. Stiefelhagen. Semi-supervised learning with constraints for person identification in multimedia data. In *IEEE International conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2013.

[4] H. Bilen, V. P. Namboodiri, and L. J. Van Gool. Object and action classification with latent window parameters. *International Journal of Computer Vision*, 106(3):237–251, 2014.

[5] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[6] M. Fischer, H. K. Ekenel, and R. Stiefelhagen. Person re-identification in tv series using robust face recognition and user feedback. *Multimedia Tools Appl.*, 55(1):83–104, Oct. 2011.

[7] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[8] S. Gong, M. Cristani, S. Yan, and C. C. Loy. *Person Re-Identification*. Springer, 2014.

[9] M. Guillaumin, J. Verbeek, and C. Schmid. Is that you? metric learning approaches for face identification. In *International Conference in Computer Vision (ICCV)*, Kyoto, Japan, Sept. 2009.

[10] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[11] C. Liu, C. C. Loy, S. Gong, and G. Wang. Pop: Person re-identification post-rank optimisation. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2013.

[12] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. IEEE, 1999.

[13] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 1, page 5, 2010.

[14] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. *arXiv preprint arXiv:1403.6382*, 2014.

[15] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. 2013.

[16] G. Shakhnarovich, L. Lee, and T. Darrell. Integrated face and gait recognition from multiple views. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–439. IEEE, 2001.

[17] M. Tapaswi, M. Beuml, and R. Stiefelhagen. Knock! knock!, who is it?: Probabilistic person identification in tv-series. In *IEEE International conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2012.

[18] P. A. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[19] C.-N. J. Yu and T. Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 1169–1176, New York, NY, USA, 2009. ACM.

[20] W.-S. Zheng, S. Gong, and T. Xiang. Reidentification by relative distance comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(3):653–668, 2013.