Accepted in IEEE Int. Conference on Advanced Video and Signal based Surveillance Workshops, 2015

Ensembles of Exemplar-SVMs for Video Face Recognition from a Single Sample Per Person

Saman Bashbaghi, Eric Granger, Robert Sabourin
Laboratoire d'imagerie de vision et d'intelligence artificielle
École de technologie supérieure, Université du Québec, Montréal, Canada
bashbaghi@livia.etsmtl.ca, {eric.granger, robert.sabourin}@etsmtl.ca

Guillaume-Alexandre Bilodeau
LITIV Lab
Polytechnique Montréal, Montréal, Canada
gabilodeau@polymtl.ca

Abstract

Recognizing the face of target individuals in a watch-list is among the most challenging applications in video surveillance, especially when enrollment is based on one reference still facial image. Besides the limited representativeness of facial models used for matching, the appearance of faces captured in videos varies due to changes in illumination, pose, scales, etc., and to camera inter-operability. A multi-classifier system is proposed in this paper for robust still-to-video face recognition (FR) based on multiple diverse face representations. An individual-specific ensemble of exemplar-SVMs (e-SVMs) classifiers is assigned to each target person, where each classifier is trained using a high-quality reference face still versus many lower-quality faces of non-target individuals captured in videos. Diverse face representations are generated from different patches isolated in facial images and face descriptors that are robust to various nuisance factors (e.g., illumination and pose) commonly encountered in surveillance environments. Discriminant feature subsets, training samples, and ensemble fusion functions are selected using faces of non-target individuals captured in videos of the scene. Experiments on videos from the Chokepoint dataset reveal that the proposed ensemble of e-SVMs outperforms state-of-the-art FR systems specialized for the single sample per person problem.

1. Introduction

Systems designed for FR in video surveillance aim to detect the presence of individuals of interest by comparing the faces captured over a network of cameras against facial models of target individuals enrolled to the system [9, 16]. In watch-list screening applications, the number of representative reference stills per target individuals is very limited [4]. Face models for matching are often designed a priori using a single high-quality reference face image captured under controlled conditions. This challenging problem is referred to as a "single sample per person" (SSPP) problem [18]. Moreover, regions of interest (ROIs) isolated within reference stills may differ significantly from those captured in operational videos, due to camera inter-operability, and to the different capture conditions. The appearance of faces captured in video under semi- or uncontrolled conditions may vary considerably according to several nuisance factors, including ambient illumination, pose, scale, expression, occlusion, and blur [3].

Different techniques have been proposed to address a SSPP problem, such as exploiting multiple face descriptors, synthetic face generation through morphing or 3D reconstruction, and using auxiliary sets to enlarge the training set [15, 18]. Recently, a system with multi-manifold learning of discriminative features from patches has been proposed to perform manifold-manifold matching [13]. Although sparse representation based classification (SRC) methods have shown a prominent performance in FR [20], they are not directly applicable to SSPP problems. To address this problem, a generic auxiliary training set has been exploited in extended SRC (ESRC) [6] to enhance the intra-class variation in order to appropriately discriminate between the probe and gallery samples. Similarly, a generic training set has been integrated with the gallery set to develop a sparse variation dictionary learning (SVDL), where an adaptive projection is jointly learned to connect the generic set to the gallery set, and to construct a sparse dictionary with sufficient variations of representations [21].

Despite improvements achieved by the abovementioned methods to handle the SSPP problem, they are not fully-adapted for still-to-video FR systems w.r.t. the following drawbacks. First, they are relatively sensitive to variations in capture conditions (e.g., considerable changes in illumination, pose, and specially occlusion). Second, samples in the generic training set are not similar to the samples in the gallery set due to the different cameras. Hence, the intra-class variation of training set may not translate to discriminative information regarding samples in the gallery set. Third, they may suffer from a high computational complex-

ity, because of the sparse coding and the large and redundant dictionaries [6, 21].

Few specialized classification systems have been proposed for still-to-video FR. Although only one high-quality reference still is available per target individual, video sequences from other non-target individuals may also be used to overcome the aforementioned challenges [4, 8, 17]. Moreover, ensemble methods have been shown to provide a high-level of performance when training data is limited and imbalanced [11]. Ensembles of template matchers based on multiple diverse face representations of a single target ROI pattern have been shown to significantly improve on the overall performance of a basic still-to-video FR system [4].

In this paper, a multi-classifier system is proposed for robust still-to-video FR, where the single reference still of a target individual is modeled using an individual-specific ensemble of exemplar-SVM (e-SVM) classifiers [14]. E-SVMs are 2-class classifiers trained using a single target facial ROI versus many non-target ROIs captured in videos from the same camera view-point. This specialized ensemble of e-SVMs models the variability in facial appearances by generating diverse face representations that are robust to different nuisance factors frequently observed in surveillance environments. Multiple face representations are generated from different face patches and face descriptors. Indeed, the abundance of non-target faces acquired from videos of unknown people in the environment (background model) is used throughout the design process to select discriminant feature sets, scores normalization, and fusion.

The contributions of this paper can be summarized as follows. First, an individual-specific ensemble of e-SVMs is proposed to discriminate between a single high-quality target ROI and an abundance of non-target ROIs from low resolution videos of the scene. Secondly, multiple face representations are generated to provide ensemble diversity and improve robustness to various perturbation factors. The performance of the proposed system is compared to state-of-the-art systems using videos of Chokepoint dataset [19].

2. Ensembles of Exemplar-SVMs Based on Multiple Representations

The block diagram of the proposed multi-classifier system is shown in Figure 1. During enrollment, an individual-specific ensemble is designed for each target individual using multiple robust facial representations and specialized e-SVM classifiers. This ensemble is robust to variability of faces by generating diverse representations (different features extracted from patches) that address common nuisance factors. E-SVM classifiers are trained under imbalanced data distributions (a single reference face still versus an abundance of non-target faces captured from video cameras in the scene). Faces captured in videos for non-target individuals appearing in the camera viewpoint are employed during design phase (enrollment of an individual). Hence, a diverse pool of e-SVM classifiers is generated during design and then classifiers' responses are combined at the score-level during operation.

2.1. Enrollment Phase

During enrollment of a target individual, the facial model is encoded into an ensemble of e-SVMs using the ROI extracted from a single high-quality reference still. The reference still ROI is first converted to gray-scale and then a facial ROI is isolated using a face detector. Then, each ROI is scaled into a common size, aligned, and then normalized for illumination invariance. Afterwards, a pool of diverse e-SVM classifiers is generated using multiple face representations extracted from patches of the reference ROI. In particular, uniform non-overlapping patches are used to improve FR of partially occluded faces [12]. For each representation, the ROI patch patterns of the target individual is combined with the corresponding ROI patterns of non-target individuals to train e-SVMs to estimate the face model. For a system with P patches, and M feature extraction techniques, enrollment involves generating a pool of MxP e-SVMs.

2.2. Multiple Feature Extraction

Feature extraction techniques are selected based on their robustness to the nuisance factors encountered in video surveillance environments [1, 2, 7]. To that end, both LBP and LPQ extract textures, while LBP is illumination invariant and LPQ is more robust to motion blur. HOG is capable of providing a high level of discrimination on a SSPP due to its modeling of gradients with different angles and orientations. Furthermore, HOG is robust to rotation and translation. Haar wavelet transform performs accurately regarding to pose changes and partial occlusion.

2.3. Exemplar-SVM Classification

In this paper, specialized Support Vector Machine (SVM) classifiers are considered for face matching as acquired in still-to-video FR. The performance of traditional SVM classifiers declines when training data is imbalanced, as the estimated boundary is skewed to the majority class (non-target ROIs). For classification of imbalanced data sets, the SVM objective function should adapt the boundary in order to decrease the effect of class imbalance [22]. Consider a training dataset $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_l, y_l)\}$ in a binary classification problem, where $\mathbf{x}_i \in R^n$ and $y_i \in \{-1, +1\}$ represent an n-dimensional data points and the classes of these data, respectively, for $i = 1, 2, \ldots, l$. Different Error Costs (DEC) methods were proposed to mod-

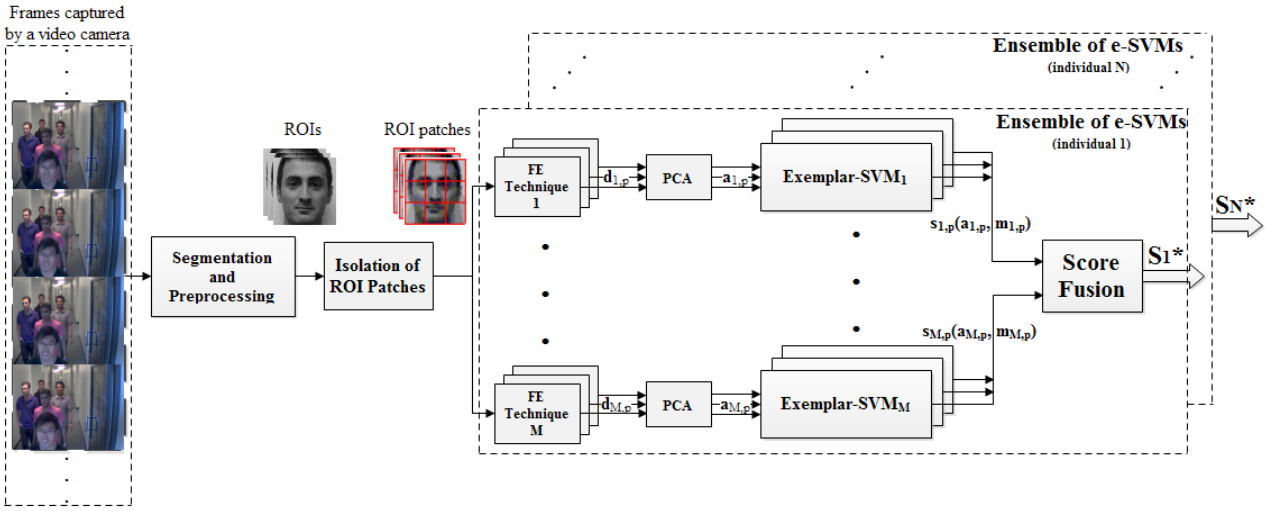

Figure 1. Block diagram of the proposed still-to-video FR system using ensemble of e-SVMs per target individual.

ify the SVM objective function, where two misclassification cost values C^+ and C^- are assigned as follows:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^2 + C^+ \sum_{[i|y_i=+1]}^{l} \xi_i + C^- \sum_{[i|y_i=-1]}^{l} \xi_i \quad (1)$$

where slack parameters ξ_i are introduced to account for misclassified examples, thus, $\sum_{i=1}^{l} \xi_i$ can be considered as a misclassification amount, and $\mathbf{w}$ is the weight vector. Constants C^+ and C^- are the misclassification costs for the positive and negative class, respectively.

The e-SVM classifier is trained using a single target sample (still ROI pattern) along with many non-target samples (ROI patterns from videos) for each individual of interest as illustrated in Figure 2. As a specialized classification approach for watch-list screening with a SSPP, training can be performed by considering all or a subset of non-target ROIs as negative samples obtained from a universal background model. Subsequently, the information of non-target individuals from the context are exploited during training to enhance the classifier generalization capability.

Let $\mathbf{a}$ be the positive sample (target ROI pattern) and U is the number of non-target (negative) samples. The formulation of the linear classifier e-SVM cost function is:

$$\min_{\mathbf{w},b} \mathbf{w}^2 + C_1 \max(0, 1 - (\mathbf{w}^T\mathbf{a} + b)) + C_2 \sum_{\mathbf{x}\in U} \max(0, 1 - (\mathbf{w}^T\mathbf{x} + b)) \quad (2)$$

where C_1 and C_2 parameters control the weight of regularization terms, and b is the bias term. Since there is

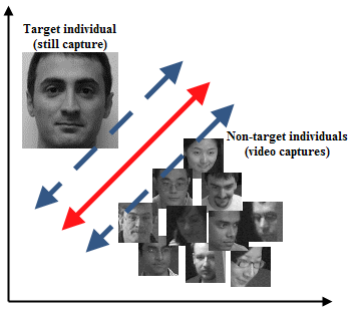

Figure 2. 2D illustration of e-SVM decision boundary for an individual of interest enrolled in the watch-list.

only one positive sample in the training set, its error is weighted much higher than the negative samples. The calibrated score of e-SVM for the given ROI $\mathbf{a}$ and the learned regression parameters (α_a,β_a) is computed as follows [14]:

$$f(\mathbf{x}|\mathbf{w},\alpha_a,\beta_a) = \frac{1}{1+e^{-\alpha_a(\mathbf{w}_a^T-\beta_a)}} \quad (3)$$

E-SVMs possess some potential benefits in designing individual-specific classifier systems with multiple face representations. The number of non-targets appears to provide enough constraints to the SSPP problem. The amount of non-targets cannot affect the accuracy of decision boundary due to estimate support vectors that are highly similar to each target [14]. Hence, it can be applied suitably even for large databases containing a few exemplars in the training set, e.g., as acquired in the watch-list screening.

Since each e-SVM is highly specialized to the target individual, the largest margin (decision boundary) will be obtained by training under imbalanced data exploiting differ-

ent regularization parameters, where it provides more freedom in defining the decision boundary. Therefore, it is less sensitive to class imbalance than other classification techniques, such as neural network and decision tree [22]. E-SVM as a passive learning approach impose no extra training overhead and compensate the imbalance data in the optimization process. Combining e-SVMs into an ensemble may prevent over-fitting issue and simultaneously provide higher generalization [10]. This method can be also interpreted as an approach to order and to select the representative non-targets by visual similarity to the target individuals.

2.4. Operational Phase

During operation, several people may appear in video frames as illustrated in Figure 1, while some of them are considered as individuals of interest. Thus, segmentation and preprocessing step are performed on each frame in order to capture face(s), and then the resulting ROI is scaled into a common size. Afterward, multiple face descriptors i = 1, 2, ..., M are extracted from each patch p = 1, 2, ..., P. PCA is employed to either rank features or to reduce dimension of face descriptors. Thereafter, each e-SVM classifier provides a classification score $\mathbf{S_{i,p}}(a_{i,p})$ between every patch ROI pattern $a_{i,p}$ and the corresponding patch model m_{i,p^j} (classifier parameters trained and preserved for each specific patch), where j=1, 2, ..., N indicates the number of individuals of interest. Classifiers scores are finally fed into the score fusion module after score normalization to obtain the final score $\mathbf{S_j^*}$.

Fusion at score-level among face patches and descriptors are applied on the ensembles to achieve higher accuracy and robustness, as follows: (1) score-level fusion of patches attempts to combine the scores generated among patches using multiple classifiers trained per each patch, and (2) score-level fusion of descriptors within the ensemble to provide the final score. In the former fusion strategy, P classifiers are used, while PxM classifiers are exploited in the latter.

3. Experimental Results

Chokepoint video dataset[1] [19] is selected as a benchmark for large-scale FR based on its characteristics to simulate real-world scenarios, especially in watch-list application. In this paper, different aspects of the proposed framework are evaluated experimentally using Chokepoint. First, experiments assess the performance of classifiers trained on ROI patterns extracted using different feature extraction techniques. Second, experiments investigate the impact of using patch configurations on the performance. Finally, the performance of score-level fusion of classifiers within the ensembles are compared.

[1] http://arma.sourceforge.net/chokepoint/

3.1. Methodology for Validation

To constitute the watch-list, 5 high-quality still ROIs of individuals of interest as shown in Figure 3 are selected randomly with neutral expression. Random examples of ROIs captured from videos are also illustrated in Figure 3. Videos of 10 unknown people that are assumed as background (to exploit a global view of the scene) are employed during enrollment phase. Thus, the rest of videos including 10 other non-target individuals are associated for the testing process along with videos of 5 watch-list individuals. Therefore, target individuals (one at a time) and unknown individuals within the test videos participate in each test iteration.

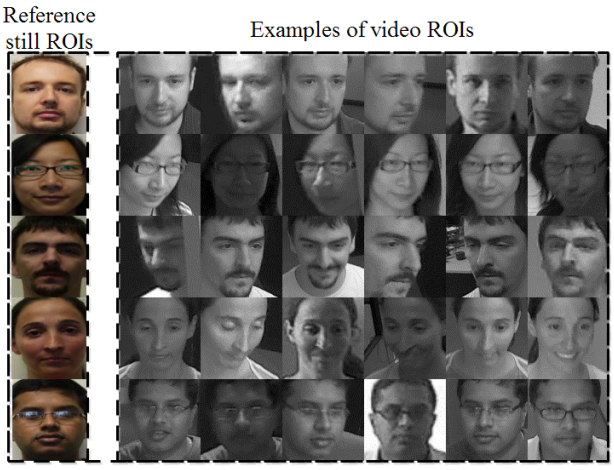

Figure 3. Illustration of ROIs captured from 'neutral' reference stills of 5 target individuals, as well as, random examples of their ROIs captured from video sequences in the Chokepoint dataset.

Libsvm [5] is used in order to train each exemplar SVM. The same regularization parameters $C_1 = 1$ and $C_2 = 0.01$ for all exemplars ($\mathbf{w}$ of a target sample is 100 times greater than non-targets) are chosen based on the imbalance ratio. Furthermore, the size of the reference stills and captured ROIs are scaled to 48x48 pixels. It should be noted that the features extracted prior to the training phase must be normalized between 0 and 1, as well as, output scores of classifiers. Hence, min-max normalization is used in this regard using non-target faces.

Ensemble of template matchers (TMs) [4] and SVDL [21] are considered as state-of-the-art systems to compare with the proposed system. In SVDL experiment, 5 high-quality stills belonging to individuals of interest are considered as a gallery set and low-quality videos of non-target individuals are employed as a generic training set to learn a sparse variation dictionary. Three regularization parameters λ_1, λ_2, and λ_3 set to 0.001, 0.01, and 0.0001, respectively according to the default values defined in SVDL. The number of dictionary atoms are initialized to 80 based on the number of stills in the gallery set, where it is a trade-off be-

Table 1. Average pAUC(20%) and AUPR performance of the proposed system over all Chokepoint videos using patches and 4 different feature extraction techniques.

ROI - Patch Configurations	Face Representations							
	LBP (max: 59)		LPQ (max: 256)		HOG (max: 500)		Haar (max: 2304)	
	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR	pAUC	AUPR
1 (48x48 pixels) block	77.86±2.53	72.12±7.18	77.93±1.80	69.13±7.10	**86.08±1.70**	**81.71±6.34**	71.12±3.08	67.54±8.92
4 (24x24 pixels) blocks	79.53±2.34	74.71±8.76	79.2±2.66	76.65±8.40	**91.03±0.84**	**88.02±4.32**	84.41±2.38	81.82±7.42
9 (16x16 pixels)blocks	81.68±2.04	77.38±6.37	85.03±1.12	82.18±6.90	**98.44±0.78**	**96.64±2.12**	82.50±1.16	80.46±6.20

tween the computational complexity and the level of sparsity. Additionally, 100 dimensional Eigenvectors are computed using the pixel intensities of faces as a feature set.

The performance of watch-list screening systems are evaluated at the transaction-level by the Receiver Operating Characteristic (ROC) curve. A global scalar metric of the detection performance is the Area Under ROC curve (AUC), which can be interpreted as the probability of classification. Precision-recall (PR) curve can also estimate the performance considering the target individuals under imbalanced data situation. Recall can be defined as TPR and precision (PR) is computed as follows $\mathrm{PR} = \frac{\mathrm{TP}}{\mathrm{TP+FP}}$. In transaction-level analysis, system performance are provided using partial AUC (pAUC) and Area Under Precision-Recall (AUPR). The AUPR is suitable to illustrate the global accuracy of the system in the skewed imbalanced data circumstances. To achieve statistically significant results, the experiments are iterated 5 trials for different groups of 5 individuals of interest, and then the average values of pAUC and AUPR for all individuals in the watch-list are reported along with standard deviations.

3.2. Results and Discussion

Performance of the proposed watch-list screening system is evaluated with different feature extraction techniques, where score-level fusion among patches and descriptors are considered using 1, 4, and 9 blocks (48x48, 24x24, and 16x16 pixels, respectively). It is worth noting that the output scores of classifiers trained over each patch are combined to provide the final score for each representation using averaging. Since the dimension of these representations are inconsistent and due to complexity and to avoid over-fitting, their dimensions are reduced using PCA[2]. The average values of pAUC(20%) and AUPR along with standard errors are presented in the Table 1 for different face descriptors and patch configurations.

As shown in Table 1, using score-level fusion of patch-based method with 9 blocks mostly outperforms the performance of without using patch (1 block) and 4 blocks. In terms of comparing among feature extraction techniques, HOG provides better performance. The maximum dimension of features that each feature extraction technique produces is also mentioned, while PCA is employed to reduce and rank them. It can be concluded that training a separate classifier for each patch provides higher performance, contrary to training one classifier on the one block. Particularly, it is confirmed that the better facial model the e-SVM trained on, the higher performance achieved. Since each face descriptors performs inconstantly, applying fusion among them can essentially capture their advantages.

In order to evaluate the proposed system against the state-of-the-art systems, SVDL [21] and ensemble of template matchers (TMs) [4] are considered. The performance of applying fusion to combine the descriptors within the ensemble at feature-level (concatenation) and score-level to combine classifiers scores (mean function) is presented in Table 2 versus the state-of-the-art systems.

Table 2. Average pAUC(20%) and AUPR performance of the proposed system over all Chokepoint videos using score-level fusion of descriptors within ensembles against state-of-the-art systems.

Systems / Performance	pAUC	AUPR
SVDL [21]	47.70±1.20	40.14±4.12
Ensemble of TMs [4]	85.60±1.40	82.78±7.06
Ensemble of e-SVMs (1 block)	92.28±0.54	90.95±2.84
Ensemble of e-SVMs (4 blocks)	98.58±0.40	97.34±1.82
Ensemble of e-SVMs (9 blocks)	**100±0.00**	**99.24±0.38**

Table 2 shows that using fusion of descriptors within the ensemble significantly outperforms feature extraction techniques individually either with or without patches at transaction-level compare to Table 1. Accordingly, it can be also concluded that exploiting patch-based method appropriately along with accurate e-SVM classifiers trained within the ensembles lead to a robust face screening system, where patches' sizes 16x16 pixels perform better than 24x24 and 48x48 pixels. Hence, the bigger the pool of e-SVMs, the more robust the system reached overall.

As shown in Table 2, ensemble of e-SVMs greatly outperforms ensemble of TMs and SVDL. Performance of the screening system using SVDL is quite poor, mostly because of the remarkable differences between target face stills in the gallery set and video faces in the generic training set

[2]For PCA projection, the first 64 eigenvectors are selected as feature sets for LPQ, HOG and Haar descriptors.

in terms of quality and appearances, as well as, lots of non-targets (unseen individuals that are not enrolled in the gallery) observed during operation. Since each faces captured should be assigned to one of the target still in the gallery, therefore, many false positive will trivially occur. It is worth mentioning that SVDL can only apply as a global N-class classifier with a higher time complexity in contrast to the proposed ensemble of 2-class e-SVM classifiers, specifically due to sparse optimization and classification during operational phase.

4. Conclusion

This paper presents a robust multi-classifier system for still-to-video FR that is specialized for watch-list screening applications with a SSPP. Several feature extraction techniques and local patches are employed to generate a diversified ensemble of exemplar-SVM per individual. Feature extraction techniques are chosen precisely based on their robustness against variety of nuisance factors encountered occasionally in VS environments. Accordingly, results confirm that using multiple robust face representations for design of facial models and encoding them into an ensemble of adapted classifiers favorably achieve an accurate system. Meanwhile, employing representative non-target samples is required to optimize performance due to estimate the classifiers parameters, discriminant feature selection, and ensemble fusion functions. Simulation results indicate that training a separate classifier for each patch and combining their scores outperforms a single classifier trained using a long feature vector of concatenated patches. Consequently, score-level fusion of descriptors within the ensemble provides significantly better performance compared to the state-of-the-art systems.

Acknowledgment

This work was supported by the Fonds de Recherche du Québec - Nature et Technologies.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face description with local binary patterns: Application to face recognition. *PAMI, IEEE Trans on*, 28(12):2037–2041, 2006. 2

[2] T. Ahonen, E. Rahtu, V. Ojansivu, and J. Heikkila. Recognition of blurred faces using local phase quantization. In *ICPR*, pages 1–4. IEEE, 2008. 2

[3] J. R. Barr, K. W. Bowyer, P. J. Flynn, and S. Biswas. Face recognition from video: A review. *IJPRAI*, 26(05), 2012. 1

[4] S. Bashbaghi, E. Granger, R. Sabourin, and G.-A. Bilodeau. Watch-list screening using ensembles based on multiple face representations. In *ICPR*, pages 4489–4494, 2014. 1, 2, 4, 5

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011. 4

[6] W. Deng, J. Hu, and J. Guo. Extended src: Undersampled face recognition via intraclass variant dictionary. *PAMI, IEEE Trans on*, 34(9):1864–1870, 2012. 1, 2

[7] O. Déniz, G. Bueno, J. Salido, and F. De la Torre. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 32(12):1598–1603, 2011. 2

[8] Z. Huang, S. Shan, H. Zhang, S. Lao, A. Kuerban, and X. Chen. Benchmarking still-to-video face recognition via partial and local linear discriminant analysis on cox-s2v dataset. In *ACCV*, pages 589–600. 2013. 2

[9] M. D. la Torre, E. Granger, P. V. Radtke, R. Sabourin, and D. O. Gorodnichy. Partially-supervised learning from facial trajectories for face recognition in video surveillance. *Information Fusion*, 24(0):31–53, 2015. 1

[10] Q. Li, B. Yang, Y. Li, N. Deng, and L. Jing. Constructing support vector machine ensemble with segmentation for imbalanced datasets. *Neural Computing and Applications*, 22(1):249–256, 2013. 4

[11] Y. Li, W. Shen, X. Shi, and Z. Zhang. Ensemble of randomized linear discriminant analysis for face recognition with single sample per person. In *Automatic Face and Gesture Recognition (FG),*, pages 1–8. IEEE, 2013. 2

[12] S. Liao, A. K. Jain, and S. Z. Li. Partial face recognition: Alignment-free approach. *PAMI, IEEE Trans on*, 35(5):1193–1205, 2013. 2

[13] J. Lu, Y.-P. Tan, and G. Wang. Discriminative multimanifold analysis for face recognition from a single training sample per person. *PAMI, IEEE Trans on*, 35(1):39–51, 2013. 1

[14] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, pages 89–96. IEEE, 2011. 2, 3

[15] F. Mokhayeri, E. Granger, and G.-A. Bilodeau. Synthetic face generation under various operatinnal conditions in video surveillance. In *ICIP*, 2015. 1

[16] C. Pagano, E. Granger, R. Sabourin, G. Marcialis, and F. Roli. Adaptive ensembles for face recognition in changing video surveillance environments. *Information Sciences*, 286:75–101, 2014. 1

[17] C. Shaokang, M. Sandra, H. Mehrtash T, S. Conrad, B. Abbas, L. Brian C, et al. Face recognition from still images to video sequences: A local-feature-based framework. *EURASIP on image and video processing*, 2011. 2

[18] X. Tan, S. Chen, Z.-H. Zhou, and F. Zhang. Face recognition from a single image per person: A survey. *Pattern Recognition*, 39(9):1725–1745, 2006. 1

[19] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *CVPRW*, pages 74–81. IEEE, 2011. 2, 4

[20] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *PAMI, IEEE Trans on*, 31(2):210–227, 2009. 1

[21] M. Yang, L. Van Gool, and L. Zhang. Sparse variation dictionary learning for face recognition with a single training sample per person. In *ICCV*, pages 689–696, 2013. 1, 2, 4, 5

[22] Z.-Q. Zeng and J. Gao. Improving svm classification with imbalance data set. In *NIPS*, pages 389–398, 2009. 2, 4