Computer Vision & Machine Learning

Alexandros Iosifidis
@
Department of Electrical and Computer Engineering
Aarhus University

Content

Here we:
- We will briefly describe what is a (scientific) report
- We will give some guidelines on how to write such a report
- We will follow the ways proposed/used in some of the most popular scientific conferences/journals of <u>Machine Learning and Computer Vision</u>

Keep in mind that:
- The way that a report is written varies in different disciplines, depending on the type of information it should convey to the reader
- Very similar content can be written in different ways, depending on the templates/norms followed by different conferences and journals. This means that what is important is the information we want to convey and not the exact specific way that the paper is structured!

AARHUS
UNIVERSITET

Department of Electrical and
Computer Engineering

Computer Vision &
Machine Learning

What is a paper/report?

A commonly used structure:

- Title

- Abstract

- Introduction

- Theoretical Analysis – Method description

- Experiments and results

- Discussion (can be part of experimental section)

- Conclusions and Summary

- Appendix (many times it is omitted)

- References

What is a paper/report?

Examples:

What is a paper/report?

Examples:

Title, authors and affiliations

Parts of a paper

Abstract:

- Gives a summary of the whole report
- Outlines -purpose, research method,
  findings, main conclusions and
  recommendations
- Mainly past tense
- Usually written last!

Keywords:

- Buzzwords summarizing the fields of work
- Usually used for searching/grouping
  papers
- Usually 3-8 keywords

Class-Specific Kernel Discriminant Analysis Revisited: Further Analysis and Extensions

Alexandros Iosifidis, *Senior Member, IEEE*, and Moncef Gabbouj, *Fellow, IEEE*

Abstract—In this paper, we revisit class-specific kernel discriminant analysis (KDA) formulation, which has been applied in various problems, such as human face verification and human action recognition. We show that the original optimization problem solved for the determination of class-specific discriminant projections is equivalent to a low-rank kernel regression (LRKR) problem using training data-independent target vectors. In addition, we show that the regularized version of class-specific KDA is equivalent to a regularized LRKR problem, exploiting the same targets. This analysis allows us to devise a novel fast solution. Furthermore, we devise novel incremental, approximate and deep (hierarchical) variants. The proposed methods are tested in human facial image and action video verification problems, where their effectiveness and efficiency is shown.

Index Terms—Approximation, class-specific kernel discriminant analysis (CSKDA), incremental learning, low-rank kernel regression (LRKR), regularization.

I. INTRODUCTION

SUBSPACE learning techniques have an important role in statistical machine learning, due to the well-known *curse of dimensionality* [1], which states that when the data dimensionality increases, the volume of the data representation space increases and the available (training) data become sparse. In such high-dimensional spaces, learning statistical models requires the utilization of an enormous amount of training data. The most well-known and commonly employed subspace learning technique is principal component analysis (PCA) [2], which defines a subspace for data projection that preserves most of the available information (in the sense of minimal l_2 norm-based reconstruction error). While PCA is able to determine the directions of the highest variance of the original high-dimensional feature space, it is an unsupervised technique and it is not able to increase class discrimination.

When the objective is to determine a subspace that enhances class discrimination, supervised techniques exploiting the labeling information available for the training data

re employed. Linear discriminant analysis (LDA) [3]–[8] is one of the most well-studied techniques and has been found very effective in many applications [9]–[12]. By assuming unimodal classes following normal distributions, the optimal discriminant subspace is obtained by maximizing the so-called Fisher ratio which is defined on the between-class and within-class scatter matrices. Several variants that overcome the unimodality assumption of LDA have also been proposed [13]–[16]. In addition, regression models exploiting the within-class and between-class scatter defined in LDA have been recently proposed [17]. Another important limitation of LDA is the fact that the maximal subspace dimensionality is restricted by the number of classes C forming the problem at hand. This is due to the fact that the rank of the between-class scatter matrix is at most equal to $C-1$. A direct consequence of this is that in binary (two-class) problems, LDA and its variants are able to determine a subspace formed by only one dimension, which might not be the optimal choice for class discrimination.

In order to overcome the latter limitation of LDA, class-specific approaches have been proposed [18]–[22]. In class-specific subspace learning techniques, the objective is to determine the optimal feature space highlighting the discrimination between a given class of interest (noted as positive class hereafter) from the remaining world (i.e., the data not belonging to the positive class, forming the hereafter called negative class). An advantage of this approach is that, by defining appropriate out-of-class and intraclass scatter matrices, the maximal dimensionality of the learned subspace is not restricted by the number of classes, but can be up to the number of samples forming the class under consideration and, thus, feature spaces of higher dimensionality (when compared to LDA) can be determined. This in turn leads to better class discrimination and better performance [19], [21], [22].

For the cases where nonlinear discrimination criteria lead to better performance, both multiclass and class-specific subspace learning techniques can be extended to their nonlinear counterparts by exploiting the well-known *kernel trick* [23]. Kernel extensions exploit the so-called *kernel function* $\kappa(\cdot, \cdot)$ defined on pairs of D-dimensional data $\{x_i, x_j\}$ and expressing dot products of the data representations in the so-called *feature space* $\mathcal{F}$, i.e., $\kappa(x_i, x_j) \triangleq \phi(x_i)^T\phi(x_j)$, where $\phi(\cdot)$ is a nonlinear function such that $x_i \in \mathbb{R}^D \to \phi(x_i) \in \mathcal{F}$. The application of a linear subspace learning technique in $\mathcal{F}$ corresponds to
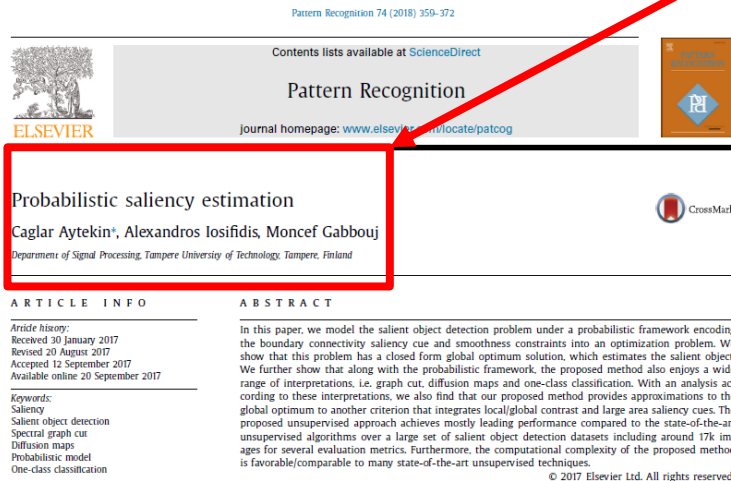
Parts of a paper

Introduction:
- Outlines context, background and purpose of the work
- Defines terms and sets limits of the research
- The reader/audience can easily identify what, how, why
- Provides literature review in text form, along with limitations of the existing works that the current work addresses
- Mainly uses past tense and can be written later although presented first

Parts of a paper

Problem statement:

- In case of methodological/theoretical work, it can be used to describe the problem that is targeted in the work
- It provides notations and definitions that will be used later in the paper
- It can be omitted (but the definitions and notations will be in any case provided in later sections)

Parts of a paper

Related work:

- It provides a detailed description of the related techniques/methods that have been proposed/used until now

- It is used to give to the reader a relatively good knowledge on the problem and highlight which are the specific roles that necessitate the existence of the current work

- It can be mathematical (describing the math behind the existing solutions) when the work is technical/theoretical

- It can be text describing existing approaches/solutions when it is a methodological work. In such cases it can be omitted, if this discussion is included in the introduction.

Parts of a paper

Methodology:

- Describes how the work was done. It can contain:

  - methodology description (e.g. description of different blocks of a pipeline)
  - mathematical description of the new model, along with derivations/proofs, etc.
  - Analysis of the complexity/cost of the model/ method.

Parts of a paper

Methodology:

- Describes how the work was done. It can contain:
  - methodology description (e.g. description of different blocks of a pipeline)
  - mathematical description of the new model, along with derivations/proofs, etc.
  - algorithmic description/pseudocode, etc.
  - discussion on the properties of the model/method, and possibly extensions/simplifications

AARHUS
UNIVERSITET

Department of Electrical and
Computer Engineering

Computer Vision &
Machine Learning

Parts of a paper

Experiments:
- It provides the description of experiments
  conducted in order to test the performance
  of the current model/method
- It provides comparisons of the current model/
  method with existing (related) methods/models
- Comparisons need to be both in terms of
  performance and efficiency (when necessary)

Parts of a paper

Experiments:

- It provides the description of experiments conducted in order to test the performance of the current model/method
- It provides comparisons of the current model/ method with existing (related) methods/models
- Comparisons need to be both in terms of performance and efficiency (when necessary)
- It should clearly define the experimental proto-col, the hyper-parameter selection process, and the competing methods/models

Parts of a paper

Experiments:

- It provides the description of experiments conducted in order to test the performance of the current model/method

- It provides comparisons of the current model/method with existing (related) methods/models

- Comparisons need to be both in terms of performance and efficiency (when necessary)

- It should clearly define the experimental protocol, the hyper-parameter selection process, and the competing methods/models

- Data description (and availability to the reader)

- Any pre-processing steps applied need to be adequately described

Parts of a paper

Experiments:

- RESULTS!

- Comparisons (quantitative and qualitative)

- Observations and possible reasons

Parts of a paper

Experiments:
- RESULTS!
- Comparisons (quantitative and qualitative)
- Observations and possible reasons
- Illustrations of performance (depending on the metrics used, e.g. Precision-Recall, ROC)
- Qualitative illustrations of results

Fig. 3. Precision–recall curves for the proposed method PSE, QCUT [20], RBD [15], ST [41], DSR [10], MC [16] and GP [36]. (Top row, from left to right) Curves for datasets MSRA10k, DUTOMRON, ECSSD. (Bottom row, from left to right) Curves for datasets JUDD and SED2.

Table 1
Maximum F_β measures of the proposed PSE and the unsupervised state-of-the art methods. Top, second and third performing methods are indicated by red, blue and green colors respectively.

	PSE	QCUT	GP	RBD	ST	DSR	MC
MSRA10k	0.8797	0.8720	0.8610	0.8559	0.8669	0.8346	0.8476
DutOmron	0.6705	0.6821	0.5838	0.6305	0.6296	0.6265	0.6274
ECSSD	0.7825	0.7778	0.7457	0.7181	0.7504	0.7371	0.7419
JUDD	0.5034	0.5093	0.4519	0.4590	0.4571	0.4558	0.4623
SED2	0.8444	0.8233	0.7704	0.8302	0.8149	0.7900	0.7713

Table 2
Performance of the compared methods with respect to AUC measures. Top, second and third performing methods are highlighted in red, blue and green colors, respectively.

	PSE	QCUT	GP	RBD	ST	DSR	MC
MSRA10k	0.9581	0.9240	0.9579	0.9489	0.9557	0.9528	0.9451
DutOmron	0.9010	0.8664	0.8672	0.8895	0.8913	0.8945	0.8827
ECSSD	0.9205	0.8836	0.9109	0.8896	0.9092	0.9089	0.9051
JUDD	0.8344	0.7740	0.7889	0.8177	0.7964	0.8186	0.8160
SED2	0.8925	0.8269	0.8717	0.8695	0.8956	0.8922	0.8549

A good method should perform well with respect to most if not all evaluation metrics. Fig. 4 provides ROC curves for the proposed method against the competing ones. The proposed method PSE is the top performing method in terms of ROC curves, except for SED2 dataset where PSE gives very high recall performance in low false positive rates. Note that one of the leading methods, in precision–recall curves, QCUT does not perform well with respect to ROC curves. The area under the curve (AUC) measures are provided in Table 2.

From Table 2, we observe that PSE is leading in all datasets except for SED2 where it trails the leading method ST by a small margin. GP, DSR and ST also perform well in AUC measures (mostly 2nd and 3rd places); however, they are not as successful in precision–recall curves and F-measure.

Finally, we compare the performance of the compared methods in terms of mean squared errors in Table 3. PSE is the top perform-

Table 3
Performance of the compared methods in terms of MSE errors. Top, second and third performing methods are highlighted in red, blue and green colors, respectively.

	PSE	QCUT	GP	RBD	ST	DSR	MC
MSRA10k	0.0588	0.1064	0.0606	0.0647	0.0568	0.0779	0.0690
DutOmron	0.0843	0.0869	0.1197	0.0833	0.0911	0.0848	0.0905
ECSSD	0.0948	0.1488	0.1056	0.1156	0.0984	0.1188	0.1067
JUDD	0.1264	0.1356	0.1579	0.1484	0.1370	0.1347	0.1296
SED2	0.0923	0.1456	0.1146	0.1047	0.0929	0.1144	0.1147

AARHUS
UNIVERSITET

Department of Electrical and
Computer Engineering

Computer Vision &
Machine Learning

Parts of a paper

Experiments:
 - RESULTS!

 - Comparisons (quantitative and qualitative)

 - Observations and possible reasons

 - Illustrations of performance (depending on
   the metrics used, e.g. Precision-Recall, ROC)

 - Qualitative illustrations of results

Fig. 2. Qualitative comparison of unsupervised methods. (From left to right: Image (IM), and the saliency maps obtained by DSR [10], MC [16], RBD [15], ST [41], GP [36], QCUT [20] and our PSE, and the ground truth (GT).
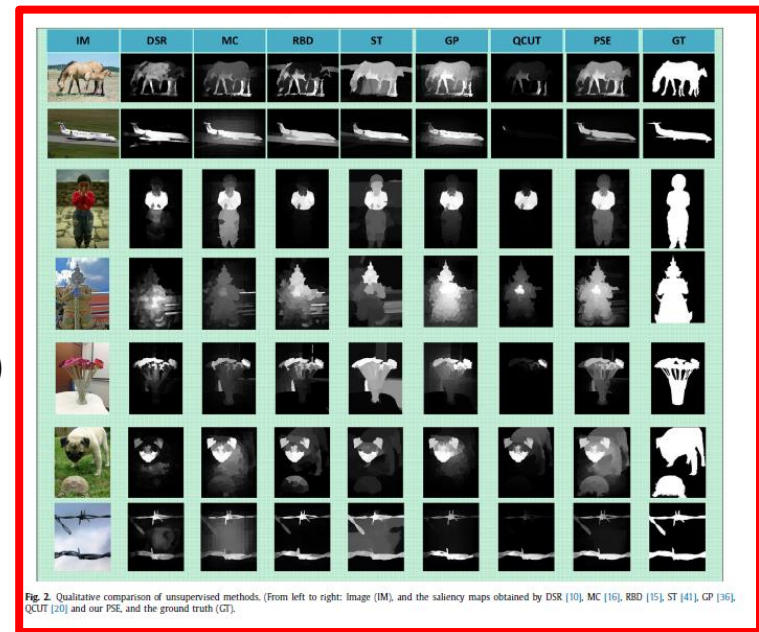
3.4. Comparison with the unsupervised state-of-the-art

We compare the proposed PSE method with the top-5 performing unsupervised techniques, according to a recent study [40], namely QCUT [20], RBD [15], ST [41], DSR [10] and MC [16]. DRFI [42] was also ranked as one of the top methods in the study; however, it is a supervised method and, hence, it is not included in our comparisons. The saliency maps for the above algorithms were downloaded from the website [30] related to the study in [40]. We further include comparisons with a recently proposed method GP [36]. The saliency maps for GP were extracted using the code provided by the authors.

First, we provide a visual comparison of the competing algorithms in Fig. 2. The images contain relatively simple images (rows 1,2), an object with a few visually distinct parts (row 3,5), an object that is highly textured (row 4) and objects touching the image boundary (row 6,7). In all images, we observe superior performance of PSE when compared to the other methods.

In Fig. 3, we provide precision–recall curves for the unsupervised state-of-the-art methods in the datasets mentioned above. One may quickly observe that although some methods such as

ST perform well in relatively simpler datasets, such as SED2 and MSRA10k, they fail to provide a good performance in more complex datasets, such as JUDD and DutOmron. One of the top performing methods QCut tends to produce saliency maps that have relatively low precision on higher recalls when compared to other methods. This is evident from the quick *fall* of the curve in this region. Furthermore, due to its hard boundary background assumption in SED2 dataset, this method fails to preserve its leading performance. It is worth noting that the proposed PSE method is consistently one of the top two performing methods in precision–recall curves. PSE does not suffer from the sharp drop in precision at high recall as QCUT. Furthermore, it handles objects touching the image boundary as observed from the plots of SED2 dataset.

In Table 1, we provide maximum F_β measures for $\beta^2 = 0.3$. In three of the datasets (i.e. MSRA10k, ECSSD and SED2) PSE gives the top performance and on DutOmron and JUDD datasets, it achieves second best performance with a relatively small gap with the leading method.

Since their evaluation is based on different criteria, it is highly likely that one may observe a different leaderboard in ROC curves and AUC measures than precision–recall curves and F-measures.

Parts of a paper

Conclusions:
- Summary of the work
- Key observations
- Limitations and possible extensions

Parts of a paper

Appendices:

- In case an in-depth analysis, mathematical proof needs to be included for completeness of the work it is placed here
- It is used in order to make the 'flow' of the paper easier

Parts of a paper

References:

- List of prior work

LRKRR problem in (30) is able to address singularity issues and has been proven to provide more robust solutions.

Following an analysis similar to that of Section IV-A, we restrict the projection matrix $\mathbf{W}$ to be a low-rank matrix, i.e., $\mathbf{W} = \mathbf{QB}$, where $\mathbf{B} \in \mathbb{R}^{d \times d}$ and $\mathbf{Q} \in \mathbb{R}^{|\mathcal{F}| \times d}$. By doing so, we restrict the rank of the projection matrix to be at most equal to $\min(d, |\mathcal{F}|)$. Then, using $\mathbf{Q} = \mathbf{\Phi A}$, we have

$$\hat{\mathcal{J}}_{\text{LRKRR}} = \left\| \mathbf{B}^T(\mathbf{Q}^T\mathbf{\Phi}) - \mathbf{T} \right\|_F^2 + C\|\mathbf{QB}\|_F^2$$
$$= \left\| \mathbf{B}^T(\mathbf{A}^T\mathbf{K}) - \mathbf{T} \right\|_F^2 + C\|\mathbf{\Phi AB}\|_F^2 \quad (31)$$

where $\mathbf{A} \in \mathbb{R}^{N \times d}$. By determining the saddle point of $\hat{\mathcal{J}}$ with respect to $\mathbf{B}$, we obtain $\mathbf{B} = (\mathbf{A}^T(\mathbf{KK}+C\mathbf{K})\mathbf{A})^{-1}\mathbf{A}^T\mathbf{KT}^T$. By substituting $\mathbf{B}$ in (18), we obtain

$$\hat{\mathcal{J}}_{\text{LRKRR}} = \left\| \mathbf{TKA}(\mathbf{A}^T(\mathbf{KK}+C\mathbf{K})\mathbf{A})^{-1}\mathbf{A}^T\mathbf{B} - \mathbf{T} \right\|_F^2$$
$$= \text{const.} - 2Tr\left((\mathbf{A}^T\tilde{\mathbf{S}}\mathbf{A})^{-1}\mathbf{A}^T\tilde{\mathbf{S}}_p\mathbf{A} \right) \quad (32)$$

where $\tilde{\mathbf{S}} = \mathbf{KK} + C\mathbf{K}$ and $\tilde{\mathbf{S}}_p = \mathbf{KT}^T\mathbf{TK}$. That is, $\hat{\mathcal{J}}_{\text{LRKRR}}$ is optimized by solving the generalized eigen-decomposition problem $\tilde{\mathbf{S}}_p\mathbf{a} = \lambda\tilde{\mathbf{S}}\mathbf{a}$, $\lambda \neq 0$, which is a regularized version of (16). This can be easily verified by using (1) and (21)

$$\mathbf{W}^T(\mathbf{S}+C\mathbf{I})\mathbf{W} = \mathbf{A}^T\mathbf{\Phi}^T(\mathbf{\Phi\Phi}^T+C\mathbf{I})\mathbf{\Phi A}$$
$$= \mathbf{A}^T(\mathbf{KK}+C\mathbf{K})\mathbf{A}$$
$$= \mathbf{A}^T\tilde{\mathbf{S}}\mathbf{A} \quad (33)$$

and $\mathbf{W}^T\mathbf{\Phi J}_p\mathbf{\Phi}^T\mathbf{W} = \mathbf{A}^T\mathbf{KJ}_p\mathbf{KA} = \mathbf{A}^T\mathbf{KT}^T\mathbf{TKA} = \mathbf{A}^T\tilde{\mathbf{S}}_p\mathbf{A}$.

Thus, we can conclude that the regularized CSKDA optimization problem, is equivalent to that of LRKRR using target vectors satisfying (21).

REFERENCES

[1] G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Trans. Inf. Theory*, vol. 14, no. 1, pp. 55–63, Jan. 1968.

[2] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, NY, USA: Wiley, 2000.

[3] J. Ye, "Least squares linear discriminant analysis," in *Proc. Int. Conf. Mach. Learn.*, Corvallis, OR, USA, 2007, pp. 1087–1093.

[4] T. Zhang, B. Fang, Y. Y. Tang, Z. Shang, and B. Xu, "Generalized discriminant analysis: A matrix exponential approach," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 40, no. 1, pp. 186–197, Feb. 2010.

[5] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with L1-norm," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 828–842, Jun. 2014.

[6] X. Cai, C. Ding, F. Nie, and H. Huang, "On the equivalent of low-rank linear regressions and linear discriminant analysis based regressions," in *Proc. SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Chicago, IL, USA, 2013, pp. 1124–1132.

[7] H. Wang, X. Lu, Z. Hu, and W. Zheng, "Fisher discriminant analysis with L1-norm," *IEEE Trans. Cybern.*, vol. 44, no. 6, pp. 828–842, Jun. 2014.

[8] X. Peng, J. Lu, Z. Yi, and R. Yan, "Automatic subspace learning via principal coefficients embedding," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2572306.

[9] A. Iosifidis, A. Tefas, and I. Pitas, "Activity-based person identification using fuzzy representation and discriminant learning," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 530–542, Apr. 2012.

[10] A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis," *Comput. Vis. Image Understand.*, vol. 116, no. 3, pp. 347–360, 2012.

[11] C.-X. Ren, Z. Lei, D.-Q. Dai, and S. Z. Li, "Enhanced local gradient order features and discriminant analysis for face recognition," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2015.2484356.

[12] F. Dornaika and A. Bosaghzadeh, "Exponential local discriminant embedding and its application to face recognition," *IEEE Trans. Cybern.*, vol. 43, no. 3, pp. 921–934, Jun. 2013.

[13] M. Zhu and A. M. Martinez, "Subclass discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 8, pp. 1274–1286, Aug. 2006.

[14] A. Iosifidis, A. Tefas, and I. Pitas, "On the optimal class representation in linear discriminant analysis," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 9, pp. 1491–1497, Sep. 2013.

[15] A. Iosifidis, A. Tefas, and I. Pitas, "Kernel reference discriminant analysis," *Pattern Recognit. Lett.*, vol. 49, pp. 85–91, Nov. 2014.

[16] A. Maronidis, A. Tefas, and I. Pitas, "Subclass graph embedding and a marginal fisher analysis paradigm," *Pattern Recognit.*, vol. 48, no. 12, pp. 4024–4035, 2015.

[17] A. Iosifidis, A. Tefas, and I. Pitas, "Graph embedded extreme learning machine," *IEEE Trans. Cybern.*, vol. 46, no. 1, pp. 311–324, Jan. 2016.

[18] Y. P. Li, J. Kittler, and J. Matas, "Face verification using client specific fisher faces," in *Proc. Int. Conf. Stat. Directions Shapes Images*, Leeds, U.K., 2000, pp. 63–66.

[19] G. Goudelis, S. Zafeiriou, A. Tefas, and I. Pitas, "Class-specific kernel-discriminant analysis for face verification," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 570–587, Sep. 2007.

[20] S. Zafeiriou, G. Tzimiropoulos, M. Petrou, and T. Stathaki, "Regularized kernel discriminant analysis with a robust kernel for face recognition and verification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 526–534, Mar. 2012.

[21] S. R. Arashloo and J. Kittler, "Class-specific kernel fusion of multiple descriptors for face verification using multiscale binarised statistical image features," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 12, pp. 2100–2109, Dec. 2014.

[22] A. Iosifidis, A. Tefas, and I. Pitas, "Class-specific reference discriminant analysis with application in human behavior analysis," *IEEE Trans. Human-Mach. Syst.*, vol. 45, no. 3, pp. 315–326, Jun. 2015.

[23] B. Scholkopf and A. J. Smola, *Learning With Kernels*. Cambridge, MA, USA: MIT Press, 2001.

[24] D. Cai, X. He, and J. Han, "Speed up kernel discriminant analysis," *Int. J. Very Large Data Bases*, vol. 20, no. 1, pp. 21–33, 2011.

[25] C. K. I. Williams and M. W. Seeger, "Using the Nyström method to speed up kernel machines," in *Proc. Adv. Neural Inf. Process. Syst.*, Denver, CO, USA, 2001, pp. 682–688.

[26] P. Drineas, R. Kannan, and M. Mahoney, "Scalable kernel clustering: Approximate kernel K-means," *arXiv:1402.3849v1*, pp. 1–15, 2014.

[27] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, 2007, pp. 1177–1184.

[28] N. Pham and R. Pagh, "Fast and scalable polynomial kernels via explicit feature maps," in *Proc. SIGKDD Int. Conf. Knowl. Discov. Data Min.*, Chicago, IL, USA, 2013, pp. 239–247.

[29] A. Iosifidis, M. Gabbouj, and P. Pekki, "Class-specific nonlinear projections using class-specific kernel spaces," in *Proc. IEEE Int. Conf. Big Data Sci. Eng.*, Helsinki, Finland, 2015, pp. 17–24.

[30] A. Iosifidis and M. Gabbouj, "Scaling up class-specific kernel discriminant analysis for large-scale face verification," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 11, pp. 2453–2465, Nov. 2016, doi: 10.1109/TIFS.2016.2582562.

[31] S. Pang, S. Ozawa, and N. Kasabov, "Incremental linear discriminant analysis for classification of data streams," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 35, no. 5, pp. 905–914, Oct. 2005.

[32] H. Zhao and P. C. Yuen, "Incremental linear discriminant analysis for face recognition," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 38, no. 1, pp. 210–221, Feb. 2008.

[33] D. Chu, L.-Z. Liao, M. K.-P. Ng, and X. Wang, "Incremental linear discriminant analysis: A fast algorithm and comparisons," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 11, pp. 2716–2735, Nov. 2015.

[34] N. Kwak, "Implementing kernel methods incrementally by incremental nonlinear projection trick," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2565683.

[35] W. W. Y. Ng, X. Tian, Y. Lv, D. S. Yeung, and W. Pedrycz, "Incremental hashing for semantic image retrieval in nonstationary environments," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2016.2582530.

[36] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[37] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, 2006.