Comparing interrelationships between features and embedding methods for multiple-view fusion

Roberta Piroddi, Yannis Goulermas, Simon Maskell, and Jason Ralph
School of Electrical Engineering, Electronics and Computer Science
University of Liverpool
{rpiroddi, goulerma, smaskell, jfralph,}@liverpool.ac.uk

***Abstract*—Manifold embedding techniques have properties that render them attractive candidates to learn a compact and general representation of a three dimensional spatial object. In turn this representation can be used for object recognition through classification. This paper presents a comparative study of several supervised spectral embedding techniques and their relationship with the feature space used to describe the exemplars which act as inputs to an embedding procedure. By concentrating on this aspect, we are able to highlight preferential combinations between feature description and embedding, and we formulate recommendations on the use of such methods for fusing multiple views of an object to recognize it under variable poses.**

I. Introduction

Multiple-view object recognition is a challenging task receiving increasing attention due to its industrially wide applications [1], [2]. In particular, the ability to recognize a three dimensional object under a variable set of environment appearances, and poses, resides at the core of complex systems geared towards intelligent automation [3], [4], surveillance [5], [6] and advanced computer graphics for animation [7], [8].

In computer vision and machine learning, multiple-view recognition may indicate a variety of tasks dealing with the integration of alternative object views from heterogeneous sources and sensors [9], [10], [11]. In this paper, we concentrate on a spatial interpretation of multiple view, interpreted as images taken by a unique sensor from multiple view points and viewing angles [12]. The aim is to find a representation that encapsulates robustly the salient characteristics of the observed object in a way that is invariant to the appearance modifications induced by the object pose [13].

Manifold embeddings have mathematical properties that render them amenable to be applied to such representation learning goals[14], [15]. Although manifold embeddings have been studied for dimensionality reduction and especially in the context of text analysis [16], relatively little work has been dedicated to adapt such methods to multiple view learning. This is in spite of the fact that notable recent works have demonstrated an ability to generalize to new situations from small training datasets [17], [18], [19]. There exist a large number of embedding methods to choose from. It is unclear which ones are good candidates to be used for object recognition, which characteristics make them good candidates, and indeed how to evaluate their potential usefulness for the task.

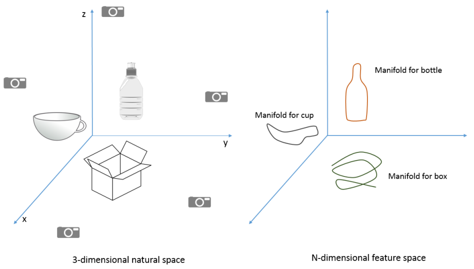


Fig. 1. Diagrammatic illustration of the mapping between the visual appearance of an object in a 3D space and its corresponding manifold in a $N-$dimensional feature space.

In this study we start addressing these questions. In particular, we concentrate on one aspect that is seldom discussed, but we found of decisive importance: the role of descriptors in the overall embedding performance. Our aim is to provide useful indications to practitioners of which combinations of features and embeddings may exhibit useful properties for tasks of multiple-view fusion and recognition. We compare six supervised spectral embedding and five popular feature descriptors. We have found distinctive outcomes that highlight a reinforcing relationship between feature spaces and their topological projections in an embedding sub-space. From these findings, we suggest effective combinations of features and embeddings as well as indicating promising expansions of this inquiry.

II. Embedding methods

Manifolds [20] are sets of points that describe the geometry or topology of an observation maintaining a local direct mapping to an Euclidean representation. Mathematically, they constitute a simpler but complete description of spatial properties of objects embedded in a more complex space [21]. Manifolds are useful representations of the appearance of object under variations that may be due to object pose and style, and environmental conditions [22], [23], [24]. The assumption enabling this use of manifolds is that the visual appearance of an object expresses itself as a set of points that form a manifold as a topological subset of a feature space, as illustrated in Figure 1.

Images acquired by sensors placed at different viewing angles capture samples of a theoretical continuous appearance manifold corresponding to an object category (left hand side of Figure 2). Manifold embeddings learn the representation

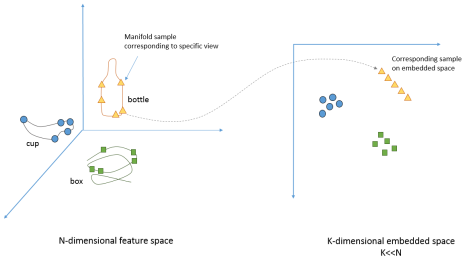

Manifold sample
corresponding to specific view
bottle
cup
box
Corresponding sample
on embedded space
N-dimensional feature space
K-dimensional embedded space
K<<N

Fig. 2. Manifold learning for object recognition is based on the assumption that different views of the object exist as samples of a continuous manifold defined on the feature space.

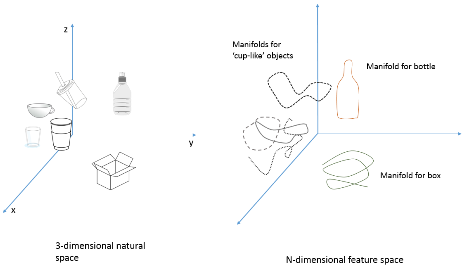


Fig. 3. A manifold embedding is said to exhibit the ability to generalize if similarly shaped objects are also nearest neighbors in the embedded space.

of a manifold by using an unorganized set of samples [25]. In practical terms, embedding is an optimization procedure that searches for a lower dimensional, and therefore simpler, unfolding of the manifold that is embedded in a higher dimension feature space. As such, manifold embeddings are usually employed for dimensionality reduction and complex dataset visualization. It is also easier to learn a lower dimensional representation of an object category [15]. This makes embeddings generalizable descriptions of objects to be used for recognition [18].

A learned embedding is the basis for a generalizable representation if geometrically and semantically similar objects generate manifolds that are in close proximity in a feature space chosen to describe them. As shown in Figure 3, the ideal scenario is that of closely clustered manifolds of similar objects. In this case, when samples are taken into considerations, these constitute the nearest neighbors of visually and/or topologically similar objects.

We consider a group of proximity embeddings called spectral embeddings. As it is shown in [15], these rely on the same optimization strategy to learn a manifold representation. This is the optimization of the trace of the spectrum of the data. As we cast the object recognition task as one of classification, we consider supervised embedding techniques.

Given a set of data points (or samples) $\{\mathbf{x}_i\}_{i=1}^n$ of dimension d, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{id}]^T$, we want to learn a set of optimal embeddings $\{\mathbf{z}_i\}_{i=1}^n$ of dimension k, where $k << d$. The conditions on the optimization are that the $n \times k$ feature matrix $\mathbf{Z} = [z_{i,j}]$ is an accurate description of the original $n \times d$ feature matrix $\mathbf{X} = [x_{i,j}]$. Additionally, the transformation operated by $\mathbf{Z}$ needs to support or improve discrimination between classes and in the case of supervised embeddings take into account the label information. Below we give a short description of the supervised embedding methods compared here.

A. Fisher Discriminant Analysis and Derived Methods

Fisher Discriminant Analysis [26] (FDA) is a linear embedding method. It therefore supports the projection $\mathbf{Z} = \mathbf{X}\mathbf{P}$, where $\mathbf{P}$ is a $d \times k$ matrix that expresses an additional linear constraint. This method computes the optimal projection matrix as:

$$\max_{\mathbf{P} \in R^{d \times k}} = \frac{tr[\mathbf{P}^T \mathbf{S}_b \mathbf{P}]}{tr[\mathbf{P}^T \mathbf{S}_w \mathbf{P}]}, \tag{1}$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ are the between class and within class scatter matrices respectively and tr is the trace. FDA is the most used embedding method for supervised learning.

From this method the Local Fisher Discriminant Analysis [27] (LFDA) uses the same optimization as in 1 but the matrices $\mathbf{S}_b$ and $\mathbf{S}_w$ are redefined to take into consideration local information. Marginal Fisher Analysis [28] (MFA) also considers local information like in LFDA and solves the optimization problem of 1, but the matrices $\mathbf{S}_b$ and $\mathbf{S}_w$ are redefined so that they are consider intraclass $K_1 - NNs$ nearest neighbors and interclass $K_2 - NPs$ nearest pairs.

B. Maximum Margin Criterion

The Maximum Margin Criterion [29] (MMC) optimal transformation matrix is obtained by solving the following optimization:

$$\max_{\mathbf{P} \in R^{d \times k}, \mathbf{P}^T\mathbf{P} = \mathbf{I}_{k \times k}} = tr[\mathbf{P}^T (\mathbf{S}_b - \mathbf{S}_w) \mathbf{P}]. \tag{2}$$

C. Discriminative Locality Alignment

Discriminative Locality Alignment [30] (DLA) solves the following optimization task:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{K_1} ||\mathbf{x}_i - \tilde{\mathbf{x}}_i^j||_2^2 - \lambda \sum_{i=1}^{n} \sum_{j=1}^{K_2} ||\mathbf{x}_i - \hat{\mathbf{x}}_i^j||_2^2, \tag{3}$$

where the data points $\tilde{\mathbf{x}}$ denote the intraclass $K_1 - NNs$ nearest neighbors of $\mathbf{x}_i$ and the data points $\hat{\mathbf{x}}$ the denote the interclass $K_2 NNs$ nearest neighbors of $\mathbf{x}_i$. Since the differences in Equation 3 can be seen as modified scatter matrices of between class and intra class differences, this method has a similar optimization structure as that in Equation 2.

D. Discriminant Neighborhood Embedding

Discriminant Neighborhood Embedding [31] (DNE) is a method that uses a linear projection, preserves the locality of the information and at the same time incorporates the label information. It is based on the following minimization:

$$\min_{\mathbf{Z}\in R^{n\times k}, \mathbf{Z}^T\mathbf{Z}=\mathbf{I}_{k\times k}} tr[\mathbf{Z}^T(\mathbf{I}_{n\times n} - \mathbf{W}^T)(\mathbf{I}_{n\times n} - \mathbf{W})\mathbf{Z}]. \quad (4)$$

Here $\mathbf{W}$ is a weight matrix and its elements are defined in the following way:

$$w_{i,j} = \begin{cases} +1, & \text{if } x_i \text{ and } x_j \text{ are intraclass K-NNs,} \\ -1, & \text{if } x_i \text{ and } x_j \text{ are interclass K-NNs,} \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

III. Experimental Set-Up

In this study we want to compare proximity based spectral embedding for supervised learning to find out which one is more promising for the purpose of multiple view object recognition. We compare six embedding methods: FDA, LFDA, MFA, MMC, DLA, and DNE. The questions we are asking revolve on which of the representation learning methods generates a low dimensional embedding that is capable to fuse multiple views of the same objects into clusters of points that are both compact and well separated from points belonging to different classes of objects.

We also want to find out which method captures salient geometrical or topological information so that categories of objects that are semantically similar are reflected in a closer clustering of representative points in the embedded space. As an example, in Figure 3 a coffee cup would be similar to a drinking glass and more similar to a bottle than to a box.

Finally, we wish to find out if the feature representation has an effect on the success of the embedding for recognition purposes, so we describe the object samples with multiple features. The features are inputs to an embedding method. The embedding method transforms a high dimensional representation in terms of features into a lower k-dimensional representation in an embedded space. For this study and for economy of visualization, we perform embeddings to reduce the feature dimensionality to $k = 2$.

A. Datasets

The dataset we employed in this study is the Columbia Object Image Library (COIL-20) dataset [32], which has been used for numerous studies of algorithmic performance for multiple-view recognition in the context of the PASCAL VOC challenge [33]. The COIL-20 dataset is composed of $M = 20$ object categories or classes. for each class, it contains $V = 72$ views, which are images of the object taken at a varying viewing angle. The viewing angles are equally spaced to describe a range of 360 degrees around the object and they are taken from the same plane and at the same distance from the object. The illumination conditions are the same for all the objects and there are no occluding elements, so to concentrate on the intrinsic properties of each embedding methods. We used the set of scale-normalized images of COIL-20.

B. Feature representations

Almost all studies addressing embedding methods omit the question of the relationship between how the observations are described in terms of features and the effectiveness of the embedding. We address this gap within the context of a specific recognition task. We compare a number of descriptors of choice for multiple-view and large spatial variation recognition tasks. All the descriptors used in this study are local descriptors.

Histograms of oriented gradients [34] (HOGS) are local descriptors which have been shown to be robust to photometric changes and form the basis of many successful invariant descriptors such as SIFT.

Edge orientation histograms [35] (EOH) utilize edge detection filters and have shown to outperform HOGs in some type of sensor-specific imagery.

Local Binary Patterns [36] (LBP) build a binary fingerprint of a local neighborhood by comparing pixels to their radial surrounding. It is a compact descriptor of intensity distribution.

Binary robust invariant scalable keypoints [37] (BRISK) uses a characteristic daisy pattern and performs pairwise comparisons of local intensity to build a binary descriptor of local intensity.

Speeded-up robust features [38] (SURF) is a blob detector that exploits the second order derivative of the image intensity in a localized neighborhood.

C. Evaluation criteria

For the purpose of finding a representation that fuses multiple views effectively to perform object recognition, we are looking at a clustering of embedded samples that makes classification of objects easy. Secondly, we look at clustering that captures some shape or geometric similarity between objects so that there is a smaller distance separating similar object categories.

To evaluate the *compactness* of the embedded clusters we consider the radius of the cluster, defined as the maximum distance of any sample from its class centroid. To evaluate the *separation* between clusters we consider the closest neighbor of a cluster, defined as the minimum distance between a sample and any sample from another class. Often in evaluating clustering these two criteria are fused into the Dunn index [39], which is a ratio that increases with separation and is inversely proportional to size of clusters. So a high Dunn Index indicates compact and separated clusters.

We consider the inter-class distances between embedded samples to evaluate the ability of the learning methods to group perceptually or geometrically similar object classes. This measures the ability to generalize to similar object or object of the same class but different styles.

IV. Experimental results

Overall the experimental results show that the choice of embedding method and its combination with a feature descriptor dramatically change the quality of the multiple-view fusion through embedding, when the objective of the embedding

TABLE I
COMPACTNESS: AVERAGE NORMALIZED RADIUS

method feature	HOG	EOH	LBP	BRISK	SURF
FDA	0.001	0	0.012	0	0.0012
LFDA	0.008	0.014	0.52	0.32	0.57
MFA	0	0	singular	singular	singular
MMC	0.24	0.29	0.46	0.30	0.55
DLA	0.037	0.08	singular	singular	singular
DNE	0.20	0.36	0.38	0.52	0.56

TABLE II
SEPARATION: AVERAGE CLOSEST NEIGHBOR

	HOG	EOH	LBP	BRISK	SURF
FDA	0.095	0.088	0.069	0.066	0.090
LFDA	0.088	0.073	0.020	0	0.024
MFA	0.073	0.079	singular	singular	singular
MMC	0.027	0.0019	0.011	0	0.0047
DLA	0	0	singular	singular	singular
DNE	0.019	0.0013	0.0026	0	0.0041

TABLE III
AVERAGE DUNN INDEX

	HOG	EOH	LBP	BRISK	SURF
FDA	80.38	1.7305e+04	NaN	NaN	NaN
LFDA	17.92	1.88e+03	NaN	NaN	NaN
MFA	Inf	2.073	singular	singular	singular
MMC	0.69	0.0088	NaN	NaN	NaN
DLA	0	0	singular	singular	singular
DNE	0.39	0.0053	NaN	NaN	NaN

is to simplify and obtain at the same time a robust object classification.

The results in Table I show the average radius of embedded objects and they are a measure of compactness of the cluster. They do not offer a clear picture if not associated to the figures in Table II that shows the average distance of an object centroid to the closest neighbor and are a measure of separation of clusters. These two measures in general would offer a clearer indication if combined in the Dunn Index, which is shown in Table III. We report this measure but we note that the characteristic behavior of the embedding methods make this metric less informative that for the evaluation of feature space clustering. As the embedding performs an extreme dimensionality reduction, the variability of the data is lost for many of the methods used and the samples are compressed into unique data points, in some cases merging views belonging to different objects. This is an example of over-compression of the dimensionality and it is particularly evident if the feature descriptors of LBP, BRISK, and SURF are used.

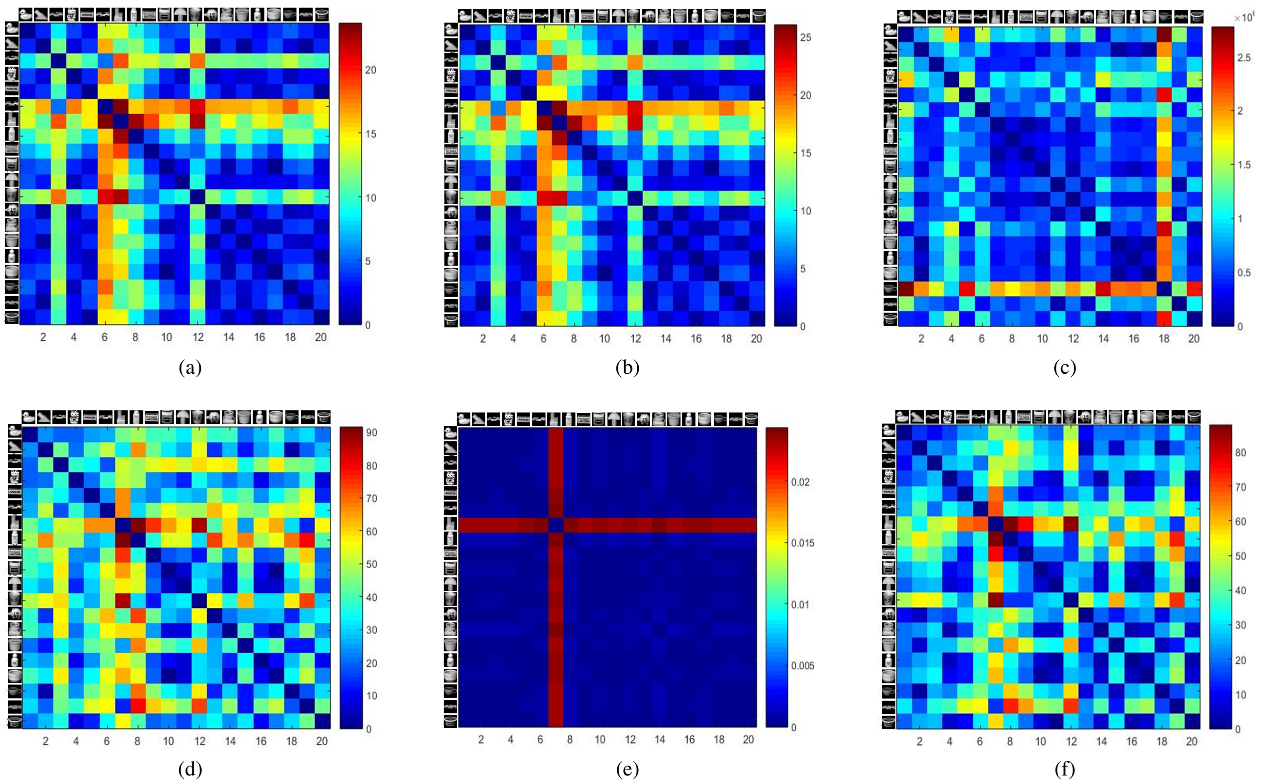
If considering the objective of obtaining a single well separated cluster per object, the methods derived originally from FDA, so FDA itself, LFDA and MFA, perform the best when combined with the feature descriptors of HOG and EOH. In general, for this dataset and feature descriptors, MFA and DLA are numerically unstable. The methods of MMC and DNE do not produce a meaningful embedding in terms of cluster description as the data points belonging to different objects are intertwined closely in the embedded space, as it is shown by inspecting Table I and II.

Another objective of the embedding for classification purposes is to cluster objects that have a similar appearance. This gives an indication of the ability of the representation to generalize. We consider the intra-class distances in the embedded space, shown in Figures 4 to 6. A representative image for each class is shown at the top and left of each figure. Figure 4 shows the results of the clustering for all embedding methods considered, using solely the HOG feature. Here FDA and LFDA contain some clusters that relate to geometrically similar objects, while another related method, MFA does not yield a useful clustering into geometrically similar categories. The clusters are in general very close together, without any distinction for shape. The MMC and DNE have some elements of useful shape clustering but clusters much more inhomogeneous that the FDA and LFDA. The DLA does not produce any useful clustering in the embedded space and this is recurring throughout the features considered. If we compare EOH shown in Figure 5 to the clustering obtained using HOGs, it is clear that while from the point of view of separating different classes both HOG and EOH performed similarly, when it comes to capture shape similarities, HOG features seem more salient for this dataset. Figure 6 compares the clustering quality for the embedding methods of FDA, LFDA, MMC, and DNE (along the rows) for the feature descriptors with which they are stable, LBP, BRISK, and SURF (down the columns). For these features, FDA performs very poorly. More surprisingly, LFDA embedding, which performed similarly to FDA for HOGs and EOHs, here obtains very different clustering results. In general, for this set of feature descriptors, LFDA, MMC and DNE obtain similar clusterings. The use of BRISK descriptor yields the less useful clustering in terms of shape similarity, while LBP and SURF perform similarly and seem able to capture some shape differentiation in terms of angular and curved surfaces.

Figure 7 gives an overview of the features embedded on the reduced dimensionality space following manifold learning. This plot shows a markedly different distribution of samples according to the combination of embedding method and feature used.

V. CONCLUSIONS AND FUTURE WORK

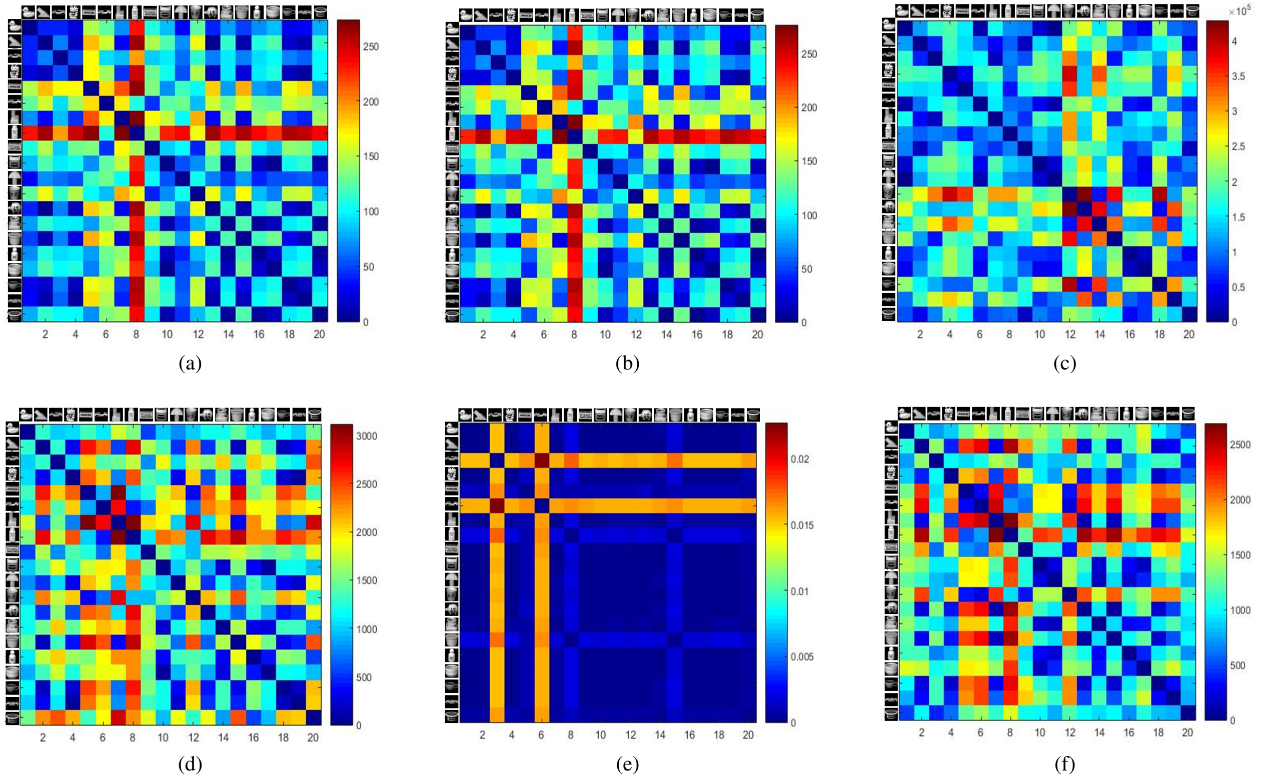
We presented a comparative study of the use of combinations of feature descriptors and manifold embedding representation learning for the purpose of multiple-view fusion. We wanted to establish which combination of embedding method and feature is promising to generalize from a relatively small amount of different views to achieve robust object recognition. While most, if not all, studies related to manifold embedding are agnostic to the feature space used as input to such methods, this study clearly demonstrates the choice of descriptor



(a) (b) (c)

(d) (e) (f)

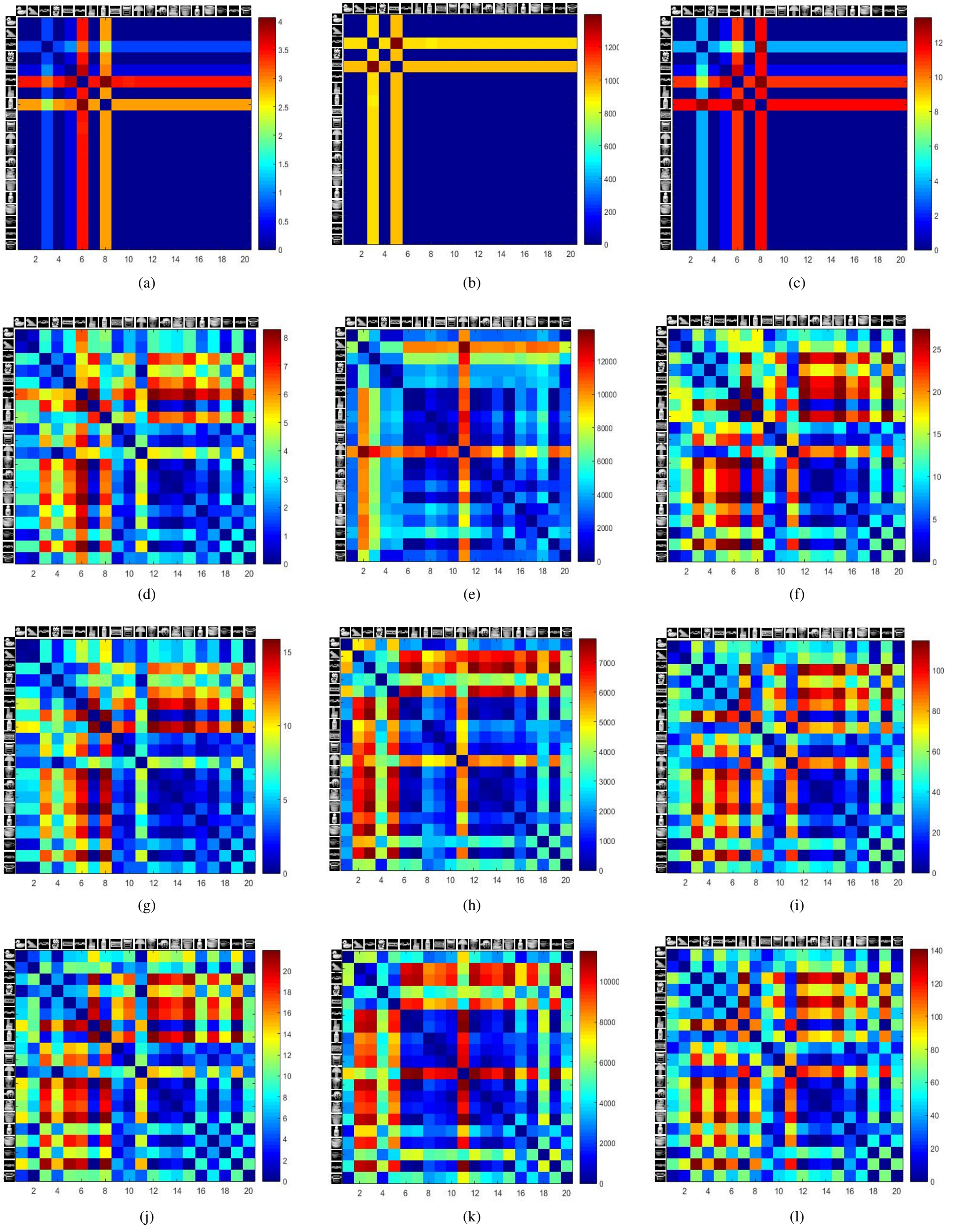
Fig. 4. Inter-object distances using HOG features for embedding methods of (a) FDA, (b) LFDA, (c) MFA, (d) MMC, (e) DLA, and (f) DNE.



(a) (b) (c)

(d) (e) (f)

Fig. 5. Inter-object distances using EOH features for embedding methods of (a) FDA, (b) LFDA, (c) MFA, (d) MMC, (e) DLA, and (f) DNE.



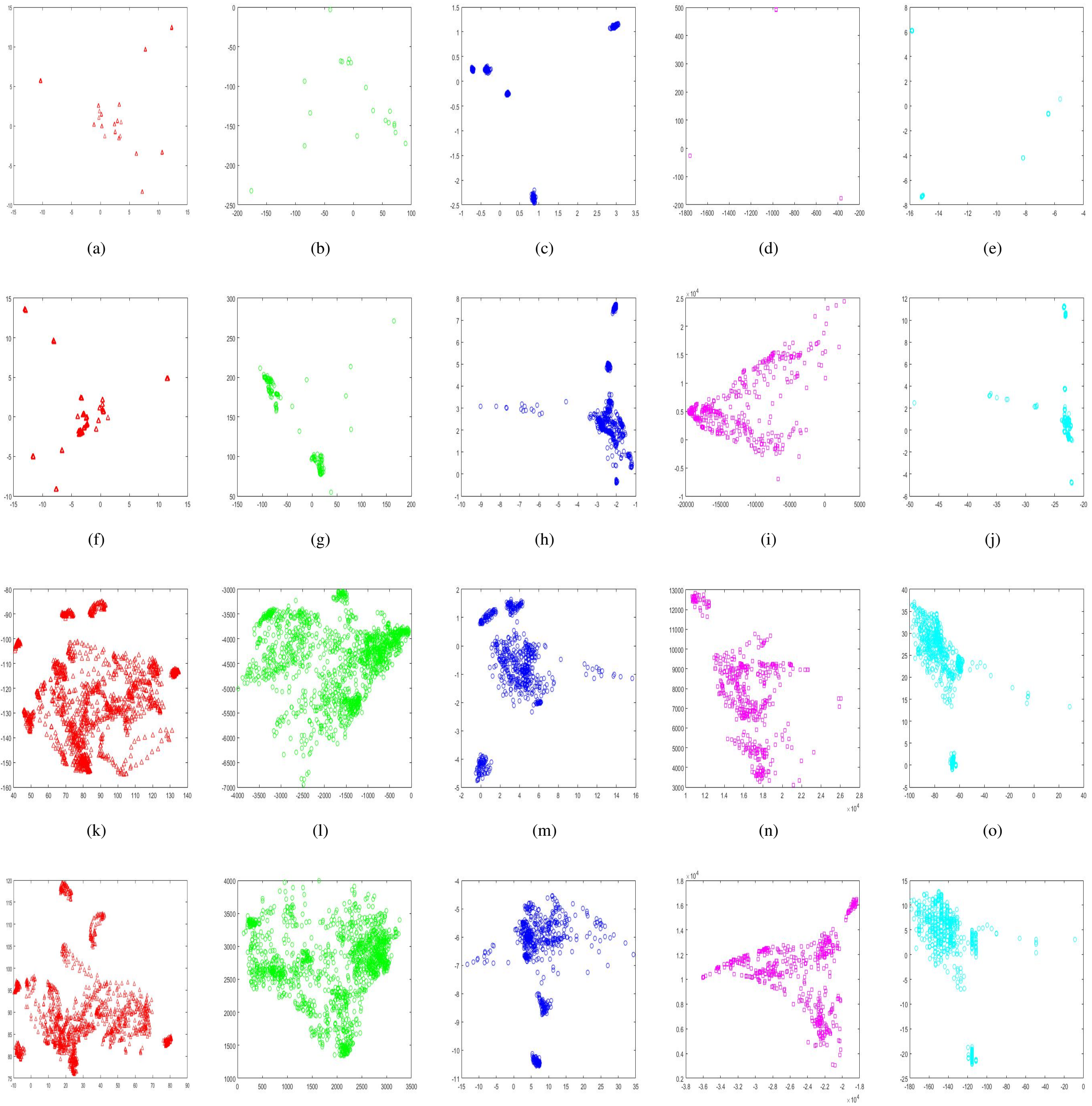
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Fig. 6. Inter-object distances using feature descriptors of LBP (in (a,d,g,j)), BRISK (in (b,e,h,k)) and SURF (in (c,f,i,l)) for embedding methods of FDA (in (a, b, c)), LFDA (in (d,e,f)), MMC (in (g, h, i)), and DNE (in (j, k, l)).



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

(k) (l) (m) (n) (o)

(p) (q) (r) (s) (t)

Fig. 7. Embedding in 2D results for COIL dataset. Rows in this grid correspond to embedding methods and columns correspond to feature descriptors. In the first row, FDA are using as inputs (a) HOGs, (b) EOHs, (c) LBP, (d) BRISK, and (e) SURF. In the second row shows LFDA (f,g,h,i,j), third row MMC (k,l,m,n,o), then DNE using same sequence of features (p,q,r,s,t).

strongly influences the performance of the embedding, at least in the case of the specific application of object recognition. For the datasets considered, we were able to recommend a combination of embedding methods and feature descriptors that perform in the most promising way and are numerically stable at the same time.

Future studies will extend this kind of analysis to more challenging datasets, in particular addressing how the invariant elements of the descriptors influence the embedding topology. Moreover, while it was interesting here to consider an uncluttered and constrained dataset, using more realistic imagery may help to identify sampling conditions that ensure better generalization. To this end, it would also be instructive to compare supervised methods to unsupervised and semi-supervised ones.

References

[1] A. Andreopoulos and J. K. Tsotsos, "50 years of object recognition: Directions forward," *Computer Vision and Image Understanding*, vol. 117, no. 8, pp. 827–891, 2013.

[2] X. Xiao, B. Javidi, M. Martinez-Corral, and A. Stern, "Advances in three-dimensional integral imaging: sensing, display, and applications," *Applied optics*, vol. 52, no. 4, pp. 546–560, 2013.

[3] D. De Gregorio, F. Tombari, and L. Di Stefano, "Robotfusion: Grasping with a robotic manipulator via multi-view reconstruction," in *European Conference on Computer Vision*. Springer, 2016, pp. 634–647.

[4] W. Mustafa, N. Pugeault, A. G. Buch, and N. Krüger, "Multi-view object instance recognition in an industrial context," *Robotica*, vol. 35, no. 2, pp. 271–292, 2017.

[5] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern recognition letters*, vol. 34, no. 1, pp. 3–19, 2013.

[6] A. Bedagkar-Gala and S. K. Shah, "A survey of approaches and trends in person re-identification," *Image and Vision Computing*, vol. 32, no. 4, pp. 270–286, 2014.

[7] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, "Performance capture from sparse multi-view video," *ACM Transactions on Graphics (TOG)*, vol. 27, no. 3, p. 98, 2008.

[8] J. G. Ruiz, D. A. Cook, and A. J. Levinson, "Computer animations in medical education: a critical literature review," *Medical education*, vol. 43, no. 9, pp. 838–846, 2009.

[9] S. Sun, "A survey of multi-view machine learning," *Neural Computing and Applications*, vol. 23, no. 7-8, pp. 2031–2038, 2013.

[10] B. Khaleghi, A. Khamis, F. O. Karray, and S. N. Razavi, "Multisensor data fusion: A review of the state-of-the-art," *Information Fusion*, vol. 14, no. 1, pp. 28–44, 2013.

[11] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Information Fusion*, vol. 31, pp. 100–109, 2016.

[12] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[13] M. Wang, Y. Gao, K. Lu, and Y. Rui, "View-based discriminative probabilistic modeling for 3d object retrieval and recognition," *IEEE Transactions on Image Processing*, vol. 22, no. 4, pp. 1395–1407, 2013.

[14] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[15] T. Mu, J. Y. Goulermas, J. Tsujii, and S. Ananiadou, "Proximity-based frameworks for generating embeddings from multi-output data," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2216–2232, 2012.

[16] T. Mu, J. Y. Goulermas, I. Korkontzelos, and S. Ananiadou, "Descriptive document clustering via discriminant learning in a co-embedded space of multilevel similarities," *Journal of the Association for Information Science and Technology*, vol. 67, no. 1, pp. 106–133, 2016.

[17] R. Wang and X. Chen, "Manifold discriminant analysis," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 429–436.

[18] Y. Chi, E. J. Griffith, J. Y. Goulermas, and J. F. Ralph, "Binary data embedding framework for multiclass classification," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 4, pp. 453–464, 2015.

[19] H. Zhang, T. El-Gaaly, A. Elgammal, and Z. Jiang, "Factorization of view-object manifolds for joint object recognition and pose estimation," *Computer Vision and Image Understanding*, vol. 139, pp. 89–103, 2015.

[20] R. L. Bishop and R. J. Crittenden, *Geometry of manifolds*. Academic press, 2011, vol. 15.

[21] S. Lang, *Introduction to differentiable manifolds*. Springer Science & Business Media, 2006.

[22] A. Elgammal and C.-S. Lee, "Nonlinear manifold learning for dynamic shape and dynamic appearance," *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 31–46, 2007.

[23] J. Wang, Z. Zhang, and H. Zha, "Adaptive manifold learning," in *Advances in neural information processing systems*, 2005, pp. 1473–1480.

[24] M. H. Law and A. K. Jain, "Incremental nonlinear dimensionality reduction by manifold learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 3, pp. 377–391, 2006.

[25] F. Nie, D. Xu, I. W.-H. Tsang, and C. Zhang, "Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction," *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1921–1932, 2010.

[26] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Annals of human genetics*, vol. 7, no. 2, pp. 179–188, 1936.

[27] M. Sugiyama, "Dimensionality reduction of multimodal labeled data by local fisher discriminant analysis," *Journal of machine learning research*, vol. 8, no. May, pp. 1027–1061, 2007.

[28] D. Xu, S. Yan, D. Tao, S. Lin, and H.-J. Zhang, "Marginal fisher analysis and its variants for human gait recognition and content-based image retrieval," *IEEE Transactions on Image processing*, vol. 16, no. 11, pp. 2811–2821, 2007.

[29] H. Li, T. Jiang, and K. Zhang, "Efficient and robust feature extraction by maximum margin criterion," in *Advances in neural information processing systems*, 2004, pp. 97–104.

[30] T. Zhang, D. Tao, and J. Yang, "Discriminative locality alignment," in *European conference on computer vision*. Springer, 2008, pp. 725–738.

[31] W. Zhang, X. Xue, H. Lu, and Y.-F. Guo, "Discriminant neighborhood embedding for classification," *Pattern Recognition*, vol. 39, no. 11, pp. 2240–2243, 2006.

[32] S. A. Nene, S. K. Nayar, H. Murase *et al.*, "Columbia object image library (coil-20)," 1996.

[33] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[34] Q. Zhu, M.-C. Yeh, K.-T. Cheng, and S. Avidan, "Fast human detection using a cascade of histograms of oriented gradients," in *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2. IEEE, 2006, pp. 1491–1498.

[35] B. Alefs, G. Eschemann, H. Ramoser, and C. Beleznai, "Road sign detection from edge orientation histograms," in *Intelligent Vehicles Symposium, 2007 IEEE*. IEEE, 2007, pp. 993–998.

[36] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 12, pp. 2037–2041, 2006.

[37] S. Leutenegger, M. Chli, and R. Y. Siegwart, "Brisk: Binary robust invariant scalable keypoints," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2548–2555.

[38] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-up robust features (surf)," *Computer vision and image understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[39] M. Halkidi, Y. Batistakis, and M. Vazirgiannis, "Clustering validity checking methods: part ii," *ACM Sigmod Record*, vol. 31, no. 3, pp. 19–27, 2002.