Learning Compact Discriminant Local Face Descriptor with VLAD

Cheng-Yaw Low
Yonsei University, Seoul, South Korea
E-mail: chengyawlow@yonsei.ac.kr

Abstract—**Local Binary patterns (LBP) and its extensions typify the present face descriptors due to their intrinsic capability of featuring the neighborhood changes striding over every pixel. These descriptors are usually engineered in an obsolete handcrafted manner and thus sufficient prior knowledge and expertise are necessitated to assure the recognition performance. This paper outlines an improved face descriptor to the recently proposed learning-based discriminant face descriptor (DFD), coined compact discriminant local face descriptor (CDLFD). In general, the pixel discriminant matrices (PDMs) that store the LBP-like local intensity variations are pruned onto the discriminant pixel vectors (DPVs) with respect to the DFD learned feature filters and the optimal soft-sampling matrices. Different from DFD and other analogous state of the arts that cluster the extracted features into the bag-of-word representation, CDLFD encodes the DPVs as a set of vector of locally aggregated descriptors (VLADs). The global VLAD signature, i.e., the concatenation of all local VLADs, is appropriately normalized and PCA whitened to yield the globally compact representation. The CDLFD performance is scrutinized based on the standard FERET evaluation protocol and the AR dataset. The experimental results disclose that CDLFD outperforms the handcrafted LBP variants, DFD, and other face descriptors, in terms of rank-1 recognition rate (%).**

I. INTRODUCTION

Face recognition has received enormous research attention due to its practical challenges. The key factors influencing the recognition performance are variations on expressions, poses, illuminations, aging, occlusions, etc., as the face images are typically acquired under uncontrolled environment. There are numerous feature representation alternatives, i.e., holistic and local, proposed to remedy these weighty issues. The principal component analysis (PCA) [1] and the linear discriminant analysis (LDA) [2] are the most well-known holistic feature extraction instances that project the entire image space onto the orthogonal eigenspace of reduced dimension. On the other hand, Gabor wavelets [4] and local binary patterns (LBP) [5] that capture the local structures have been proven superior to the holistic descriptors. In this paper, the LBP variants are in focus.

The LBP operator is originally proposed by Ojala et al. [7] as a means of local texture descriptor. It thresholds the n adjacent pixels with respect to the patch centroid p. For a 3×3 neighborhood, the LBP is derived as an inner product of the thresholded outputs and constant weights of 2^n as follows:

$$p = \sum_{n=0}^{7} s(i_n - i_c) \cdot 2^n \quad (1)$$

where $s(.)$ indicates a sign function resulting a 1 or 0. The 8-bit LBP codes, in this example, are regionally histogrammed as the final texture descriptors. The idea of adopting LBP in face recognition is primarily motivated by its proficiency in encoding the pixel-wise particulars, e.g., edges, lines, etc., as a composition of complementarily useful micro-patterns [5]. In addition to that, it has been evidenced to be discriminative, invariant to monotonic gray-level and illumination changes.

There are numerous LBP extensions proposed to improve the original LBP, e.g., [3], [6], [8], [9] and [10]. In general, they are skillfully handcrafted based on the prior knowledge and expertise. However, these handcrafted descriptors suffer from poor recognition performance in practice. This is mainly due to the reason that inadequate discriminative features are retrieved. Furthermore, the handcrafted LBPs are constrained by computational burdens if more neighboring variability is to be accommodated. On the other hand, for LQP proposed in [9], there is no way to learn a rich codebook that considers all unknown possibilities.

In this paper, a learning-based compact discriminant local face descriptor (CDLFD), which is an extension to DFD [12], is outlined. **Fig. 1** portrays the generic CDLFD framework and some frequently used abbreviations are listed in **Table 1**. On the whole, the CDLFC training module incorporates DFD, derived in accordance to the Fisher criterions, to cultivate the discriminant filters F and the soft-sampling matrices G from the LBP-like features stored in the pixel discriminant matrices (PDMs). The compressed DPMs, dubbed discriminant pixel vectors (DPVs) hereafter, are defined based on the learned F and G. This is followed by an unsupervised K-means learning stage to cluster the DPVs into a codebook of K-centroids for VLAD definitions. The major contributions of this work are:

1. In place of the indecisive PDM extraction mechanism proposed by DFD, where the neighboring pixels to be included for PDM formulation remains empirical, this work devises a greedy neighborhood selection strategy forcing all the local patch centroids to contribute to the PDM formulation (to be detailed in **Section III**).

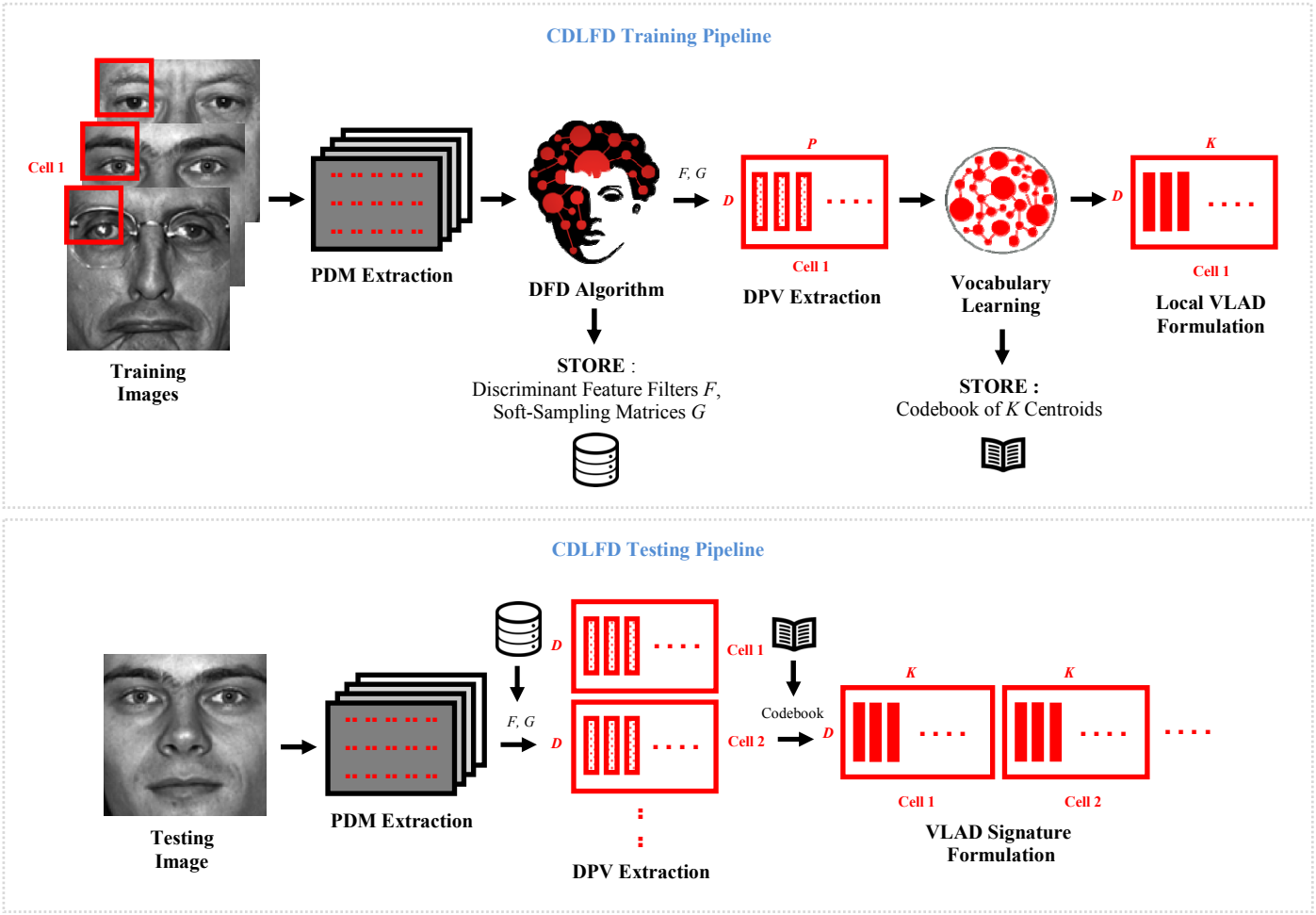


Fig. 1. The CDLFD framework constitutes separate training and testing modules. The CDLFD training module incorporates DFD algorithm to learn a unique set of discriminant feature filters F and a soft-sampling matrix G based on the PDMs extracted from each non-overlapping cell. The DPVs are formulated by retaining the most discriminative PDM entities and pruning away the least significant neighbors with respect to the cultivated F and G, respectively. A codebook is subsequently constructed via the K-means algorithm to elaborate the faces as VLAD signatures.

2. In accordance to power-law normalization, the PDMs are squashed by the signed square root (SSR) operation to alleviate the numerically dominating PDM elements, particularly at the cell boundaries padded with zeros. This stabilizes the DFD learning stage in return.
3. The DFD learned features, i.e., DPVs in this work, are congregated into multiple local VLADs of fixed-length, rather than histogramming the extracted features based on the bag-of-word (BoW) model. This is mainly due to the downside of BoW that it leads to considerable quantization loss as the BoW features only carry zero-order statistics. Other BoW demerits, e.g., the uneven histogram distribution due to burstiness, sparsity, etc., are the factors corrupting the recognition performance. The VLAD descriptor furnished with the first- and the second-order statistics is thus proposed to compensate the quantization loss.

TABLE I
FREQUENTLY USED ABBREVIATIONS

ABBREVIATION	FULL DEFINITION
PDM	Pixel Difference Matrix
DPV	Discriminant Pixel Vector
F	Discriminant Feature Filter
G	Soft Sampling Matrix
VLAD	Vector of Locally Aggregated Descriptor
CDLFD	Compact Discriminant Local Face Descriptor
DFD	Discriminant Face Descriptor (Lei's Learning Algorithm [12])

II. PRELIMINARY

This section introduces the preliminary essentials, specifically, DFD and the vector aggregation instances: BoW and VLAD, before gaining an insight into the proposed CDLFD.

A. *Discriminant Face Descriptor*

The DFD algorithm [12-14] includes a Fisher learning module based on LBP-like inputs. The data-adaptive DFD, in general, solidifies the handcrafted LBP descriptors by diminishing the within-class variability whilst magnifying the between-class margin in two perspectives:

1. The DFD discriminant feature filters F, behave as the ordinary image filters, are elicited from PDMs to single out the most representative facial traits. For the training set with N subjects, where each subject n includes m_n images, the within-class scattering matrix S_w' and the between-class scattering matrix S_b' are as follows:

$$S_w' = \sum_{n=1}^{N}\sum_{m=1}^{m_n}\sum_{p=1}^{P} F^T\left(z_{n,m}^p - \bar{z}_n^p\right)\left(z_{n,m}^p - \bar{z}_n^p\right)^T F$$

$$S_b' = \sum_{n=1}^{N}\sum_{p=1}^{P} m_n\, F^T(\bar{z}_n^p - \bar{z}^p)\,(\bar{z}_n^p - \bar{z}^p)^T\, F \tag{2}$$

   where $z_{n,m}^p \in \mathcal{R}^{D_1 \times D_2}$ is a PDM describing any pixel-p within the m-th image of the subject n; D_1 and D_2 denotes the pre-determined feature filter size and the number of neighboring pixels, respectively; $\bar{z}_n^p$ and $\bar{z}^p$ refers to the PDM mean for the subject n and the PDM grand mean for the pixel-p.

2. Instead of treating all neighboring pixels equally, DFD includes an optimal neighborhood sampling strategy to differentiate the contribution of each neighboring pixel based on the learned soft-sampling matrix G. The least significant neighbors are isolated to formulate DPVs. This problem is solved based on the Fisher's criterions, as in (3), where S_w'' and S_b'' are the within-class and the between-class scattering matrix, respectively.

$$S_w'' = \sum_{n=1}^{N}\sum_{m=1}^{m_n}\sum_{p=1}^{P}\left(z_{n,m}^p - \bar{z}_n^p\right) GG^T\left(z_{n,m}^p - \bar{z}_n^p\right)^T$$

$$S_b'' = \sum_{n=1}^{N}\sum_{p=1}^{P} m_n\,(\bar{z}_n^p - \bar{z}^p)\, GG^T(\bar{z}_n^p - \bar{z}^p)^T \tag{3}$$

Similar to 2D-LDA [22], the optimization problem of finding the optimal F and G is reformulated by combining (2) and (3), as in (4), where each F and G is fixed alternatively to obtain the eigensolution of another. The full DFD learning module is summarized in **Table II** for reference.

$$S_w = \sum_{n=1}^{N}\sum_{m=1}^{m_n}\sum_{p=1}^{P} F^T\left(z_{n,m}^p - \bar{z}_n^p\right)\, GG^T\left(z_{n,m}^p - \bar{z}_n^p\right)^T F$$

$$S_b = \sum_{n=1}^{N}\sum_{p=1}^{P} m_n\, F^T(\bar{z}_n^p - \bar{z}^p)\, GG^T(\bar{z}_n^p - \bar{z}^p)^T F \tag{4}$$

TABLE II
DFD ALGORITHM

INPUT :

A set of local PDMs extracted for a non-overlapping cell over the entire training specimens, $\{ z_{n,m}^p \in \mathcal{R}^{D_1 \times D_2}, m = 1, 2, \dots, m_n, n = 1, 2, \dots, N, p = 1, 2, \dots, P \}$, where $z_{n,m}^p$ be the PDM defined for any pixel-p in$m$-th specimen of subject n; D_1 denotes filter size; and D_2 denotes the number of neighboring pixels.

OUTPUT :

1. Discriminant Feature Filters, $F \in \mathcal{R}^{D_1 \times D_1'}$;
2. Soft Sampling Matrix, $G \in \mathcal{R}^{D_2 \times D_2'}$;

where D_1' and D_2' denotes the reduced dimension for D_1 and D_2, respectively.

INITIALIZATION :

Set $F = I^{D_1 \times D_1}$ and $G \in I^{D_2 \times D_2}$, where I is an identity matrix.

START :

FOR $t = 1, 2, \dots, T$, where T signifies the pre-defined iteration.

1. Compute within and between-class scatter matrices for F with respect to G.

$$S_w^1 = \sum_{n=1}^{N}\sum_{m=1}^{m_n}\sum_{p=1}^{P}(z_{n,m}^p - \bar{z}_n^p)\; GG^T\,(z_{n,m}^p - \bar{z}_n^p)^T$$

$$S_b^1 = \sum_{n=1}^{N}\sum_{p=1}^{P} m_n\,(z_{n,m}^p - \bar{z}^p)\; GG^T\,(z_{n,m}^p - \bar{z}^p)^T$$

   where $\bar{z}_n^p$ is the PDM mean for n-th subject and $\bar{z}^p$ is the PDM grand mean for pixel-p.

2. Solve the generalized eigenvalue problem $S_b^1 F = \lambda S_w^1 F$ to obtain the eigenvector F_0 by retaining D_1' largest eigenvalues.
3. Assign F_0 to F.
4. Compute within and between-class scatter matrices for G with respect to the F estimated in the preceding step.

$$S_w^2 = \sum_{n=1}^{N}\sum_{m=1}^{m_n}\sum_{p=1}^{P}(z_{n,m}^p - \bar{z}_n^p)^T\, FF^T(z_{n,m}^p - \bar{z}_n^p)$$

$$S_b^2 = \sum_{n=1}^{N}\sum_{p=1}^{P} m_n\,(z_{n,m}^p - \bar{z}^p)^T\; FF^T(z_{n,m}^p - \bar{z}^p)$$

5. Solve the generalized eigenvalue problem $S_b^2 G = \lambda S_w^2 G$ to obtain the eigenvector G_0 by retaining D_2' largest eigenvalues.
6. Assign G_0 to G.

END FOR

RETURN : F, G

B. *Bag-of-Word*

The BoW model, originally proposed for natural language and information retrieval, is used in computer vision to elaborate the abstracted local descriptors by a single fixed-length vector. An unsupervised clustering method, typically K-means due to

its simplicity, is employed to learn a codebook of Kcentroids (equivalent to a word dictionary) [19]. Each local descriptor is assigned to its closest codeword and the BoW representation is subsequently described as a histogram that accumulates the statistical co-occurrences of the descriptors over the codebook. It has been proven that BoW tolerates certain degree of local deformation of visual objects. One of the notorious demerits of BoW is that it merely captures the zero-order statistics, i.e., multiplicities, by discarding the important spatial traits [30]. Despite various enhancement alternatives have been proposed via the soft quantization methods [28], [29], the BoW feature remains a sparse vector of occurrence counts with artificial bursty elements and empty bins. This causes the underlying feature distribution (PDF) to be uneven and, as a consequence, corrupt the similarity scores.

C. *Vector of Locally Aggregated Descriptors*

The Fisher kernel [17] includes the first- and the second-order statistics, i.e., the approximate feature location relative to the mean and the variance in each region [19]. In general, the Fisher kernel describes how a descriptor set deviates from the average of a known parametric model, i.e., Gaussian Mixture Model (GMM) in this case, to learn a discriminative classifier. In the image representation domain, the Fisher kernel learns the latent characteristics from the extracted features to obtain a compact representation of fixed-dimension. Perronnin et al. [21] reveals that the Fisher kernel outperforms BoW under the same empirical settings. It is also found to be computationally inexpensive as it only requires a coarse codebook that ranges from 16 to 256 codewords [19]. The typical BoW descriptors necessitate a finer codebook, e.g., DFD possesses a codebook of 1000 codewords for the feature histogramming task.

VLAD is a non-probabilistic Fisher kernel proposed by Jegou et al. in [19]. The full VLAD derivation is presented in [21]. As of BoW, a data-adaptive vocabulary, i.e., a codebook C of K codewords, as defined in (6), is learned from a set of D-dimensional features V (5), the DPVs in our case, based on the K-means algorithm. For each v_p, the residual aggregating the distances between each v_p and its closest codeword μ_k is estimated as in (7):

$$V = \{ v_1, v_2, \dots, v_P \}, v_p \in \mathcal{R}^{D \times 1} \tag{5}$$

$$C = \{ \mu_1, \mu_2, \dots, \mu_K \},\ \mu_k \in \mathcal{R}^{D \times 1} \tag{6}$$

$$S_k{}' = \sum_{v_p:\, \mathrm{NN}(v_p)=k} v_p - \mu_k \tag{7}$$

where $\mathrm{NN}(v_p)$ represents the codeword index k closest to v_p. For a particular cell, the initial VLAD of $D \times K$ is vectorized to construct a local VLAD signature S of $D.K \times 1$ dimensions. For an image regionalized into $X \times Y$ non-overlapping cells, the global VLAD signature is a vector containing $D.K.X.Y$ dimensions.

To address the PDF disparity issue, the original VLAD formulation [19] outlines two normalization tricks: power-law and L2-normalization. In a nutshell, the local VLADs for each cell are normalized via the element-wise signed square root (SSR) operation, and, subsequent to that, the SSR-ed VLADs are L2-normalized. The standard VLAD formulation is listed in **Table III**.

Noting the great accomplishment of VLAD over BoW, J´egou and Chum [20] decorrelates the PCA-compressed feature through a whitening stage. In [19], to compensate the quantization losses, Jegou et al. the VLAD descriptors are yielded with respect to multiple codewords. Arandjelovic and Zisserman [18] intra-normalize each individual VLAD unit, followed by L2- normalization on the entire VLAD signature. Besides, they advocate a simple yet efficacious technique of improving the residual estimations via vocabulary adaptation for an inconsistent codebook trained on other datasets.

TABLE III
STANDARD VLAD SIGNATURE FORMULATION

INPUT :
1. A set of local descriptors extracted from a cell.
$V = \{ v_p \in \mathcal{R}^{D \times 1}, p = 1, 2, \dots, P \}$
2. A codebook of K-codeword learned via k-means clustering,
$C = \{ \mu_k \in \mathcal{R}^{D \times 1}, k = 1, 2, \dots, K \}$.
OUTPUT :
A normalized VLAD signature representing a local descriptor set,
$S \in \mathcal{R}^{D.K \times 1}$.
INITIALIZATION :
A null VLAD signature matrix, $S' = 0^{D \times K}$.
START :
FOR $p = 1, 2, \dots, P$,
1. Identify the closest codeword for each local descriptor v_n.
$k = \arg\min_l \| v_p - \mu_l \|$
2. Update the sum of residual stored in S' with respect to μ_k.
$S_k{}' := S_k{}' + v_p - \mu_k$
END FOR
3. Vectorize S' to obtain an un-normalized VLAD signature S,
where $S \in \mathcal{R}^{D.K \times 1}$.
4. Apply power law normalization to S.
$S := \mathrm{sign}(S) \times \lvert S \rvert^{\alpha},\ 0 < \alpha < 1$
5. Apply L2-normalization to S.
$S := S / \| S \|_2$
RETURN : $S \in \mathcal{R}^{D.K \times 1}$

III. COMPACT DISCRIMINANT LOCAL FACE DESCRIPTOR

In this section, the CDLFD framework that contains separate training and testing modules (as showed in **Fig. 1**) is detailed. For an image regionalized into $X \times Y$ non-overlapping cells, the input-output elements for these modules are summarized

in **Table IV**. In CDLFD, the PDMs extracted for any pixel-p residing in cell-i, denoted by $(z_{n,m}^{p})_i$, are of size $D_1 \times D_2$ each, where $D_1 = (2R+1) \times (2R+1) - 1$, $D_2 = 9$, and R refers to the LBP radius set to 2 in this paper.

TABLE IV
INPUT-OUTPUT ELEMENTS FOR CDLFD TRAINING AND TESTING MODULES

CDLFD TRAINING MODULE	
INPUT	PDMs for Training Images
OUTPUT	Discriminant Filters F, where $\mathcal{F} = \{ F_i \in \mathcal{R}^{D_1 \times D_1'}, i = 1, 2, \ldots, X \times Y \}$, $D_1 = (2R+1) \times (2R+1) - 1$, $D_1' < D_1$.
	Soft-Sampling Matrices G, where $\mathcal{G} = \{ G_i \in \mathcal{R}^{D_2 \times D_2'}, i = 1, 2, \ldots, X \times Y \}$, $D_2 = 9$, $D_2' < D_2$
	Codebook of K Codewords, where $\mathcal{C} = \{ C_i \in \mathcal{R}^{D \times K}, i = 1, 2, \ldots, X \times Y \}$. $D = D_1' \times D_2'$
	VLAD Signature S_{TR} where $S_{TR} = \{ S_i \in \mathcal{R}^{D.K \times 1}, i = 1, 2, \ldots, X \times Y \}$.
CDLFD TESTING MODULE	
INPUT	PDMs for Testing Images, α, β and γ.
OUTPUT	VLAD Signature S_{TT} where $S_{TT} = \{ S_i \in \mathcal{R}^{D.K \times 1}, i = 1, 2, \ldots, X \times Y \}$.

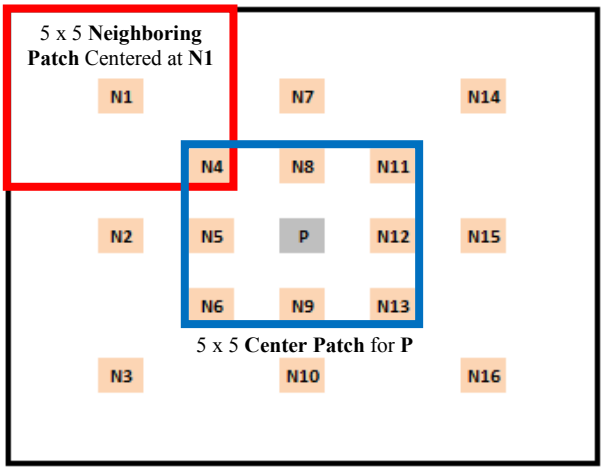
A. CDLFD Training Module

The CDLFD training module borrows the DFD algorithm to learn F_i and G_i from the LPB-like DPMs extracted from each individual cell i. The local discriminative features, i.e., DPVs, are elicited with respect to the cultivated F_i and G_i, and the K-means codebook learning phase is followed. Unlike DFD, the CDLFD training stage terminates with the VLAD formulation based on the learned codebook.

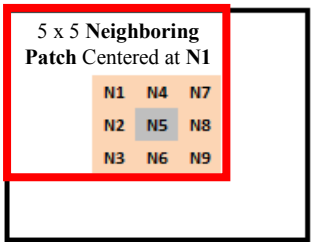
PDM Extraction

The pixel-wise PDM formulation for DFD and CDLFD are compared in **Fig. 2** and **3**. For every pixel in each cell, a chain of mini patches, following the dimensionalities of F_i and G_i, are formed within a local patch. Since these mini patches are centered at the pre-determined neighboring pixels, they are termed neighboring patches in this paper. In DFD, as depicted in **Fig. 2(a)**, 16 neighbors are selected empirically from each 13 × 13 local patch. The columnar PDMs, for R = 2, records the intensity variability between each 5 × 5 neighboring patch and the corresponding center patch for the current pixel P (see **Fig. 3(a)**). Due to the reason that this neighborhood selection mechanism is somehow subjective, it merits reassessment as it imposes a direct impact on the generalization performance.

The proposed greedy neighborhood selection is shown in **Fig. 2(b)**. In contrast to DFD, all the local patch centroids for the current pixel **P**, which overlays on **N5** in this example, are mandatorily employed for the PDM formulation. This restricts

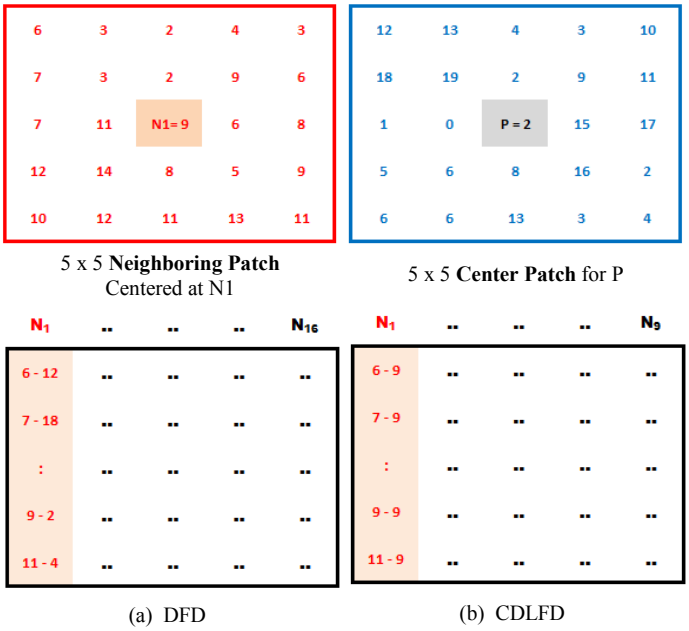


(a) DFD



(b) CDLFD.

Fig. 2. The PDM formulation for DFD (a) with respect to 16 neighboring pixels (**N1**, **N2**, …, **N16**) empirically selected from a local patch of 13 × 13, and CDLFD (b) with respect to the mandatory 9 local patch centroids (**N1**, **N2**, …, **N9**) striding over a 7 × 7 local patch.



(a) DFD (b) CDLFD

Fig. 3. The PDM formulation for DFD (a) and CDLFD (b). In DFD, the columnar PDM elements store the residuals between the 5 x 5 **neighboring patch** centered at **N1** and the **center patch** for **P**; in CDLFD, similar to LBP, the columnar PDM elements accommodate the intensity changes between the pixels within the 5 × 5 **neighboring patch** and its patch centroid **N1**.

the local patch size to 7 × 7, for R = 2, with only 9 neighbors, allowing all pixels within the local patch to be sufficiently used to construct more discriminative PDMs for the upcoming learning stage. The simplified PDM formulation in this paper, originated from [15], is illustrated in **Fig. 3(b)**. Different from DFD, each columnar PDM entry stores the residuals between the adjacent pixels of a 5 × 5 neighboring patch and its patch centroid, without requiring the intermediate central patch. The experimental results, to be presented in **Section IV**, reveals that the obligatory 9 neighbors determined from the proposed greedy selection strategy, in conjunction with the simplified PDM extraction, characterize the local texture better.

DPV Extraction

Let F_i be a set of $D_1{}'$ filters of D_1 dimensions learned for cell i, each PDM $(z_{n,m}^p)_i$ of $D_1 \times D_2$ is projected onto the feature of $D_1{}' \times D_2$ dimension. Subsequent to that, the resultant feature is pruned based on G_i by retaining the most $D_2{}'$ significant neighbors to yield a discriminant pixel matrix to be vectorized as a DPV, denoted by $(v_{n,m}^p)_i$ of $D_1' \cdot D_2'$ dimensions. The two-stage DPM to DPV projection is expressed as follows:

$$(v_{n,m}^p)_i = F_i^T \,.\, (z_{n,m}^p)_i \,.\, G_i \tag{8}$$

VLAD Formulation

VLAD approximates the GMM clustering in the Fisher kernel with non-probabilistic descriptors through K-means clustering. A complete codebook γ storing $X \times Y$ codeword sets learned based on V_i, i.e., a repository storing the DPVs extracted for cell i from all training images, is defined as:

$$\mathcal{C} = \{\, C_i \in \mathcal{R}^{D \times K}, x = 1, 2, \ldots, X \times Y \,\} \tag{9}$$

Equation (5), (6) and (7) describe how the VLAD descriptors aggregate the residuals of each DPV and its closest centroid to yield a global VLAD signature of $D \cdot K \cdot X \cdot Y$ dimensions. In addition to the SSR and L2-normalization tricks, the VLADs are defined based on multiple codewords. This is to alleviate the discrepancies of the learned vocabulary since it is usually inconsistent with the unseen testing inputs. In the experiments, the residuals obtained for each $(v_{n,m}^p)_i$ on multiple codewords are ranked and weighted accordingly by a positive constant [1, 0). This indicates that the residual for the closest codeword is to be weighted by 1, and so forth. The direct benefit of using multiple codewords will be disclosed in **Section IV**.

As the global VLAD signature is of long dimension, due to cell partitioning, it is compressed using PCA followed by a whitening step. This is to generate a compact VLAD signature representation that has been demonstrated to be fitted more to the Fisher kernel [20].

B. CDLFD Testing Module

The CDLFD testing module, in general, follows the training pipeline, except receiving the learned parameters: F_i, G_i and C_i, as inputs (see **Fig. 1**). For an unseen test image, the raw pixel-wise PDMs are SSR-normalized, projected into the expressive DPVs, and PCA whitened into a compact VLAD signature progressively. What follows is the measurement of the Cosine similarity scores over the VLAD signatures in the gallery set. The CDLFD performance assessed on the FERET and the AR datasets will be presented in the next section.

IV. EXPERIMENTS

This section scrutinizes the CDLFD performance, in terms of rank-1 recognition rate (%), in accordance with the standard FERET evaluation protocol [23] and the AR dataset [24]. The CDLFD generalization capability is compared to the leading state of the arts, and the impact analyses on VLAD variants, VLAD formulations, etc., are also provided.

A. FERET and AR

The FERET dataset is a publicly available benchmarking face repository with age, gender and ethnicity diversities. The standard FERET evaluation protocol includes a training set of 1002 frontal-view images from 429 subjects. The testing set, on the other hand, consists of a gallery set: FA with only an image per subject (1196 images in total); and 4 probe sets: FB, FC, DUP I and DUP II with facial expressions, illuminations, and time span variations (with 1195, 194, 722 and 234 images, respectively). These images are pre-processed into 128 ×128 pixels based on the annotated eye coordinates. Some FERET exemplars in the FA gallery set are displayed in **Fig. 4**.

The AR dataset contains 2600 frontal-view faces acquired from 100 subjects: 50 males and 50 females, in two occasions. Each subject provides 26 faces of various expressions: neutral, smile, anger, scream; illumination conditions: left or/and right lighting on; and object occlusions: sun-glasses and scarf. **Fig. 5** illustrates the 13 exemplars of each 165 × 120 pixels for a single subject, captured in the first contact session. In the experiments, the AR dataset is only adopted to investigate the CDLFD performance on disguises to complement the FERET protocol. It is thus separated into two groups: non-occluded and occluded. The non-occluded images, 14 for each subject, serves as inputs for the DFD training stage; and the remaining occluded faces act as probes.

Fig. 4. FERET exemplars contained in the FA gallery set.

Fig. 5. AR exemplars characterized by facial expressions, illuminations and occlusions.

B. *Parameter Configurations*

The FERET and AR training images are partitioned into 8×8 non-overlapping cells, prior to the PDM formulation step. The pixel-wise PDMs are extracted from every cell to cultivate F_i, G_i and C_i based on the parameters configured as follows: $R = 2$, $D_1 = 24$, $D_1' = 0.5 \times 24 = 12$, $D_2 = 9$, and $D_2' = 5$. **Table V** summarizes the feature dimension for the inputs and outputs involved. For K =16, the global VLAD signature is a 61,440-dimensional feature cascaded from 64 local VLADs of each 960 ($= D_1' \cdot D_2' \cdot K$) dimensions. The whitening PCA (WPCA) is applied to yield the globally compact VLAD representation of 1,000 dimensions for similarity scoring via Cosine distance.

TABLE V
PARAMETER CONFIGURATION LIST

COMPONENTS	FEATURE DIMENSION
PDM, $(\boldsymbol{z}_{n,m}^{p})_i$	24×9
$\boldsymbol{F_i}$	24×12
$\boldsymbol{G_i}$	9×5
DPV, $(\boldsymbol{v}_{n,m}^{p})_i$	$(12 \cdot 5) \times 1$
CODEBOOK, $\boldsymbol{C_i}$ ($\boldsymbol{K = 16, 32}$)	$(12 \cdot 5) \times K$
LOCAL VLAD	$12 \cdot 5 \cdot K$
GLOBAL VLAD (BEFORE WPCA)	$(12 \cdot 5 \cdot K) \cdot 64$
COMPACT VLAD (AFTER WPCA)	1000

C. *FERET : Performance Comparison with State of the Arts*

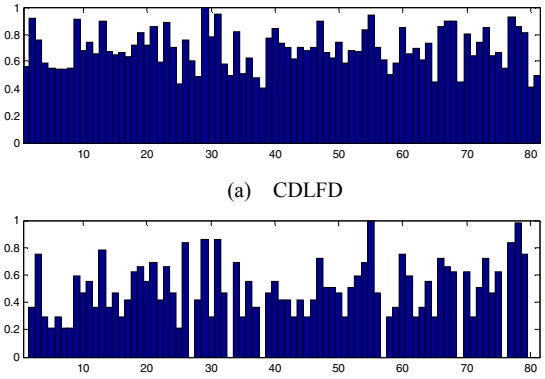
The CDLFD performance, in terms of rank-1 recognition rate (%), is compared to DFD and the influential LBP variants in this section. **Table VI** clearly discloses that the performance of CDLFD, DFD and the other two learning descriptors: DT-LBP and DLBP, on average, surpasses the handcrafted LBP, LGBPHS, LQP and POEM. These learning descriptors, on the other hand, are substantiated to be superior as LGBPHS work on the performance-guaranteed Gabor domain.

For a fair and square comparison, DFD is re-implemented in two ways based on the optimal parameters suggested: BoW with K =1000 (i.e., the original implementation), and VLAD with K =32, abbreviated as DFD_{BoW} and DFD, respectively. It is noted from **Table VI** that the DFD performance shows some inconsistencies as compared to the figures published in the original paper. This is probably due to the use of different FERET images of varying sizes and preprocessing methods. From observation, DFD and CDLFD with VLAD descriptors outperform their corresponding BoW implementation. **Fig. 6**, generated by setting K to 32, reveals that $\text{CDLFD}_{\text{BoW}}$ features, in general, are of great disparity, as the codeword occurrences only concentrate on certain bins. The bursty elements and also the empty bins would corrupt the intrinsic discriminability of the descriptor. This revelation is therefore a strong proof that CDLFD is an ideal alternative to $\text{CDLFD}_{\text{BoW}}$. In comparison to $\text{CDLFD}_{\text{BoW}}$ with K =1000, CDLFD only requires a coarse codebook with K = 32 to obtain satisfactory performance.

TABLE VI
EXPERIMENTAL RESULTS, IN TERM OF RANK-1 RECOGNITION RATE (%), FOR STATE OF THE ART FACE DESCRIPTORS BASED ON **STANDARD FERET EVALUATION PROTOCOL**

DESCRIPTORS	FB	FC	DUP I	DUP II	MEAN
* Weighted LBP [5] (TPAMI, 2006)	97.00	79.00	66.00	64.00	76.50
* LGBPHS [8] (ICCV, 2005)	98.00	97.00	74.00	71.00	85.00
* LQP [9] (BMCV, 2012)	99.80	94.30	85.50	78.60	89.55
* POEM [25] (TIP, 2012)	99.60	99.50	88.80	85.00	93.23
* DT-LBP [26] (ACCV, 2010)	99.00	100	84.00	80.00	90.75
* DLBP [27] (FG, 2011)	99.00	99.00	86.00	85.00	92.25
* DFD_{BoW} [12] (TPAMI, 2014)	99.40	100	91.89	92.30	95.88
DFD_{BoW} [12] (TPAMI, 2014)	96.07	97.94	89.20	86.75	92.49
DFD [12] (TPAMI, 2014)	97.15	98.97	91.97	88.89	94.24
$\text{CDLFD}_{\text{BoW}}$	97.91	100	89.47	87.61	93.75
CDLFD	**98.58**	**100**	**92.24**	**90.17**	**95.25**

* Note that the experimental results are extracted from the original papers.



(a) CDLFD

(b) $\text{CDLFD}_{\text{BoW}}$

Fig. 6. Feature distribution for CDLFD (a) and $\text{CDLFD}_{\text{BoW}}$ (b). It is evident that the $\text{CDLFD}_{\text{BoW}}$ features concentrate on some particular bins, where these peaks (and the empty bins) would bias the similarity scores. CDLFD, on the other hand, mitigate the BoW problems through VLAD.

D. *FERET : Impact Analysis on Different VLAD Formulations*

This section investigates the performance impact pertaining to different VLAD formulations on multiple closest codewords, by varying K_C from 1 to 5. The rank-1 recognition rates (%) listed in **Table VII** reveal that this is a practically viable trick

to fine-tune the generalization performance as the codebook is adaptively finessed to cater discrepancies on the unseen inputs.

TABLE VII
IMPACT ANALYSIS OF DIFFERENT VLAD FORMULATIONS BASED ON **STANDARD FERET EVALUATION PROTOCOL**

	K_C	FB	FC	DUP I	DUP II	MEAN
CDLFD ($K = 16$)	1	98.58	100.00	92.39	88.03	94.75
	2	98.49	100.00	93.35	88.46	95.08
	3	**98.55**	**100.00**	**93.21**	**88.89**	**95.11**
	4	98.33	100.00	92.80	88.46	94.90
	5	98.33	100.00	92.52	88.03	94.72
CDLFD ($K = 32$)	1	98.33	100.00	88.23	84.19	92.69
	2	98.58	100.00	91.27	88.89	94.71
	3	98.58	100.00	91.97	90.17	95.18
	4	**98.58**	**100.00**	**92.24**	**90.17**	**95.25**
	5	98.58	100.00	91.83	90.17	95.14

E. FERET : Impact Analysis of Intra-Normalization

The intra-normalization strategy, proposed in [18] to alleviate the burstiness problem, is examined and compared to the SSR and L2-normalization. This involves an additional step to L2-normalize every *D*-dimensional VLAD constituent, followed by SSR and L2-normalization on the local and global VLADs. **Table VIII** presents the intra-normalization impact, where **IN** is a binary operator switching the intra-normalization **ON** (1) or **OFF** (0). This normalization strategy, however, shows no improvements on the recognition rates.

TABLE VIII
IMPACT ANALYSIS OF DIFFERENT VLAD VARIANTS BASED ON **STANDARD FERET EVALUATION PROTOCOL**

	IN	FB	FC	DUP I	DUP II	MEAN
CDLFD ($K = 16$)	1	98.41	100.00	92.94	88.03	94.85
	0	**98.33**	**100.00**	**93.21**	**88.89**	**95.11**
CDLFD ($K = 32$)	1	98.66	100.00	92.24	89.89	94.95
	0	**98.58**	**100.00**	**92.24**	**90.17**	**95.25**

F. AR : Performance against Occlusions

Table IX displays the CDLFD rank-1 recognition rates (%) for VLAD defined using at most 5 closest codewords, where $K_C = 1, 2, 3, 4$ and 5, on AR images. On the whole, CDLFD evidences remarkable robustness, about 99% accuracy, to the sun-glasses and scarf disguises, under our empirical protocol. For future work, CDLFD will be investigated with respect to the protocol described in [16], where only the illuminated and the neutral expression faces, i.e., 4 images per subject, are included in the reference gallery set.

TABLE IX
EXPERIMENTAL RESULTS, IN TERM OF RANK-1 RECOGNITION RATES (%), ON **AR DATASET** AGAINST SUN-GLASSES AND SCARF OCCLUSIONS

	K_C				
	1	**2**	3	**4**	5
CDLFD ($K = 16$)	98.82	**99.16**	98.91	98.74	98.74
CDLFD ($K = 32$)	99.07	99.33	99.33	**99.41**	99.07

V. CONCLUSIONS

CDLFD is a data-driven LBP variant extending the recently proposed DFD. It borrows the DFD's idea, derived based on the Fisher's criterion, to cultivate a set of discriminant filters and soft-sampling matrices from the pixel-wise PDMs that stockpile the raw neighborhood variability. Subsequent to that, the DPMs are projected onto DPVs as the discriminant face features to be elaborated by a compact VLAD signature. The empirical results reveal that the proposed greedy neighboring selection mechanism, the SSR normalization on the PDMs, and also the whitened VLAD representation formulated based on multiple codewords are capable of improving the overall DFD performance on the standard FERET evaluation protocol. In addition, CDLFD also exhibits remarkable robustness to sunglasses and scurf disguises on the AR dataset. For future work, the performance of CDLFD will be evaluated on other face datasets of different characteristics to further validate its performance on various circumstances.

ACKNOWLEDGEMENT

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by Ministry of Science, ICT and Future Planning (2013006574) and Institute of BioMed-IT, Energy-IT and SmartIT Technology (BEST), a Brain Korea 21 Plus Program, Yonsei University.

REFERENCES

[1] M.A. Turk and A.P. Pentland, "Face Recognition Using Eigenfaces," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 586-591, June 1991.

[2] P.N. Belhumeur, J.P. Hespanha, and D.J. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, July 1997.

[3] Z. Lei, S.Z. Li, R. Chu, and X. Zhu, "Face Recognition with Local Gabor Textons," *Proc. Int'l Conf Advances in Biometrics*, pp. 49-57, 2007.

[4] C. Liu and H. Wechsler, "Gabor Feature Based Classification using the Enhanced Fisher Linear Discriminant Model for Face Recognition," *IEEE Trans. Image Processing*, vol. 11, no. 4, pp. 467-476, Apr. 2002.

[5] T. Ahonen, A. Hadid, and M. Pietikainen, "Face Description with Local Binary Patterns: Application to Face Recognition,"

IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 12, pp. 2037-2041, Dec. 2006.

[6] X. Tan and B. Triggs, "Enhanced local texture feature sets for face recognition under difficult lighting conditions," *IEEE Trans. Image Process*ing, vol. 19, no. 6, pp. 1635–1650, Jun. 2010.

[7] T. Ojala, M. Pietikainen, and D. Harwood, "A Comparative Study of Texture Measures with Classification Based on Feature Distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51-59, 1996.

[8] W. Zhang, S. Shan, W. Gao, and H. Zhang, "Local Gabor Binary Pattern Histogram Sequence (LGBPHS) : A Novel Non-Statistical Model for Face Representation and Recognition," *Proc. 10th IEEE Int'l Conf. Computer Vision*, pp. 786-791, 2005.

[9] S. ul Hussain, T. Napoleon, and F. Jurie, "Face Recognition Using Local Quantized Patterns," *Proc. British Machine Vision Conf.*, 2012.

[10] B. Zhang, S. Shan, X. Chen, and W. Gao, "Histogram of Gabor Phase Patterns (HGPP): A Novel Object Representation Approach for Face Recognition," IEEE Trans. Image Processing, vol. 16, no. 1, pp. 57-68, Jan. 2007.

[11] J. Choi, W. R. Schawartz, H. Guo and L. S. Davis, "A Complementary Local Feature Descriptor for Face Identification," *IEEE Workshop on Applications of Computer Vision*, pp. 121-128, 2012.

[12] Z. Lei, M. Pietikainen, and S. Z. Li, "Learning Discriminant Face Descriptor," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 36, no. 2, pp. 289–302, Feb. 2014.

[13] Z. Lei, D. Yi, and S.Z. Li, "Discriminant Image Filter Learning for Face Recognition with Local Binary Pattern Like Representation," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[14] Z. Lei and S.Z. Li, "Learning Discriminant Face Descriptor for Face Recognition," *Proc. Asian Conf. Computer Vision*, 2012.

[15] J. Lu, V. E. Liong, X. Zhou and J. Zhou, "Learning Compact Binary Face Descriptor for Face Recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, no. 1, pp. 1, DOI 10.1109/TPAMI.2015.2408359.

[16] T.H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma, "PCANet: A Simple Deep Learning Baseline For Image Classification?", *arXiv preprint arXiv:1404.3606*, 2014.

[17] T. Jaakkola and D. Haussler, "Exploiting Generative Models in Discriminative Classifiers," Proc. Conf. Neural Information Processing Systems, 1998.

[18] R. Arandjelovic and A. Zisserman, "All About VLAD," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1578-1585, 2013.

[19] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating Local Descriptors into a Compact Image Representation," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1704-1716, 2011.

[20] H. Jégou and O. Chum, "Negative Evidences and Co-occurences in Image Retrieval : The Benefit of PCA and Whitening," ECCV 2012, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer Berlin Heidelberg, 2012, pp. 774–787.

[21] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier, "Large-Scale Image Retrieval with Compressed Fisher Vectors," *IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[22] J. Ye, R. Janardan, and Q. Li, "Two-Dimensional Linear Discriminant Analysis," Proc. 18th Ann. Conf. Neural Information Processing Systems (NIPS), 2004.

[23] P.J. Phillips, H. Moon, S.A. Rizvi, and P.J. Rauss, "The FERET Evaluation Methodology for Face-Recognition Algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 10, pp. 1090-1101, Oct. 2000.

[24] A.M. Martinez and R. Benavente, " The AR Face Database." CVC Technical Report #24, June 1998.

[25] N.-S. Vu and A. Caplier, "Enhanced Patterns of Oriented Edge Magnitudes for Face Recognition and Image Matching," *IEEE Trans. Image Processing*, vol. 21, no. 3, pp. 1352-1365, Mar. 2012.

[26] D. Maturana, D. Mery, and A. Soto, "Face Recognition with Decision Tree-Based Local Binary Patterns," *Proc. 10th Asian Conf. Computer Vision*, pp. 618-629, 2010.

[27] D. Maturana, D. Mery, and A. Soto, "Learning Discriminative Local Binary Patterns for Face Recognition," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition and Workshops* , pp. 470- 475, 2011.

[28] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, June 2008.

[29] J. van Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. "Visual word ambiguity,", *Accepted in IEEE Trans. on Pattern Analysis and Machine Intelligence*.

[30] X. Peng, L. Wing, X. Wang, Y. Qiao, "Bag of Visual Words and Fusion Methods for Action Recognition: Comprehensive Study and Good Practice, " May 2014.