Descriptor Optimization for Multimedia Indexing and Retrieval

Bahjat Safadi, Georges Quénot

Descriptor optimization for multimedia indexing and retrieval

Bahjat Safadi and Georges Quénot
{firstname.lastname}@imag.fr
UJF-Grenoble 1 / UPMF-Grenoble 2 / Grenoble INP / CNRS, LIG UMR 5217, Grenoble, F-38041, France

***Abstract*—In this paper, we propose and evaluate a method for optimizing descriptors used for content-based multimedia indexing and retrieval. A large variety of descriptors are commonly used for this purpose. However, the most efficient ones often have characteristics preventing them to be easily used in large scale systems. They may have very high dimensionality (up to tens of thousands dimensions) and/or be suited for a distance costly to compute (e.g. χ^2). The proposed method combines a PCA-based dimensionality reduction with pre- and post-PCA non-linear transformations. The resulting transformation is globally optimized. The produced descriptors have a much lower dimensionality while performing at least as well, and often significantly better, with the Euclidean distance than the original high dimensionality descriptors with their optimal distance. The method has been validated and evaluated for a variety of descriptors using TRECVid 2010 semantic indexing task data. It has then be applied at large scale for the TRECVid 2012 semantic indexing task on tens of descriptors of various types and with initial dimensionalities from 15 up to 32,768. The same transformation can be used also for multimedia retrieval in the context of query by example and/or relevance feedback.**

I. Introduction

In multimedia indexing, a considerable research effort is directed towards the development of efficient, fast and robust indexing and retrieval systems. There are still some major challenges that need to be tackled to increase the retrieval performance of the indexing system, especially when the datasets are of large-scale. One possibility of increasing the system's performance is to carefully examine the feature normalization techniques, which have the potential to greatly decrease the error rate of the classification, and thus increase the indexing performance. However, it has been so far neglected in many research papers on multimedia indexing. In general, only a few words are devoted to the used normalization technique, even though feature normalization is a crucial step for the multimedia indexing systems.

In general, for video indexing, the Chi-square (χ^2) distance is considered to be more suitable than the Euclidean distance for comparing histogram-based visual descriptors like bag of words or BoW [1], [2]. SVM with RBF kernels can be used with both types of distance, Euclidean or χ^2. However, the χ^2 distance has two inconveniences: it is significantly more costly to compute because of the divisions in its formula and it is not compatible with PCA-based dimensionality reduction. While the Euclidean distance is conserved during the application of the PCA rotation matrix, the χ^2 distance is completely transformed, almost randomly, and might even become undefined, since it is normally computable only between vectors with positive or null components, a property which is not conserved during the application of the PCA rotation matrix.

In this paper, we investigate a simple descriptor component transformation whose goal is to make the Euclidean distance closer to the χ^2 distance. After this transformation, the Euclidean distance is expected to be as suited as the χ^2 distance for comparing histogram-based image descriptors and a SVM with a Euclidean distance-based RBF kernel is expected to be as suited as a SVM with a χ^2 distance-based RBF kernel for image classification using histogram-based image descriptors. This transformation permits a reduction of the classification time both by using a distance much simpler to compute and by being able to perform a PCA-based dimensionality reduction. We compare the classification performance on TRECVid 2010 using the multi-SVM with RBF kernels [3] with either the χ^2 or the Euclidean distance. The comparison is complicated because other and complementary normalizations can be performed either at the level of the descriptor vector (e.g. unit length normalization using either a L_1 or L_2 metric) or at the level of the descriptor components (e.g. min_max or standard deviation normalization) or a combination of several of them. Furthermore, we present an empirical evaluation of several feature normalization techniques, namely: unit length normalization (L_1 and L_2), min_max normalization, zero-mean and unit-variance normalization (σ_norm) and the power transformation. These normalization techniques are applied to several video descriptors and evaluated on the semantic indexing task of the TRECVid 2010 collection.

Another objective, of this paper, is to show that after an appropriate power transformation, the Euclidean distance becomes as effective as the χ^2 distance for image classification using SVMs with RBF kernels. Moreover, PCA-based dimensionality reduction permits a further efficiency while still being effective as well. Finally, post-PCA power transformation are also evaluated and showed to bring a further performance improvement.

II. Related work

A. Descriptor normalization methods

The main goal of descriptor or feature normalization, is to independently normalize the feature components, in such a way that their values lie within a similar range (e.g. $[0, 1]$ range). The normalization is often done by either using the component values of each vector independently (e.g. the L_1 and L_2 normalization), or by normalizing the vectors using their bin values (e.g. min_max normalization). However, other normalization techniques work directly on the values independently, and it does not consider any of the other related values (e.g. power transformation).

Let X be the set of n feature vectors of d-dimensions (components) to be normalized. In the following, we consider five techniques for feature normalization or transformation which are widely used for image and video representation. If we consider $X = (x_{ij})$ as a $d \times n$ matrix of which columns

are the descriptor vectors and rows correspond to vector components, the first two operate on columns independently, the next two operates on rows independently, and the last one operates on elements independently.

L_1 or L_2 unit length normalization: These two normalization methods scale independently the components for each vector so that the vector length becomes 1 considering either the L_1 or L_2 metric. The L_1 and L_2 normalization methods are widely used to normalize the feature vectors based on histograms including bag of words (BoW) [1], [2].

Min_max normalization: This function aims to scale the values for each feature bin (in a low-level description), so that they all fall in the range of *Lower to Upper* bounds (l, u). This normalization is used for instance in libsvm [4] with $l = 0$ and $u = 1$.

Zero-mean and unit-variance normalization (σ_norm): The feature values are normalized by subtracting the mean value μ_i for each feature bin and by dividing the result by the variance σ_i of the feature bin.

Power transformation The goal of the power transformation is to normalize the distributions of the values, especially in the case of histogram components. It simply consists of applying an $x \leftarrow sign(x) \times |x|^{\alpha}$, transformation on all components individually. The power transformation was applied by [5], in which the authors applied the power only on the *Fisher kernel* descriptor. They empirically observed that this step consistently improves the quality of the representation. They gave several complementary interpretations that justify this transform. First, it reduces the influence of bursty visual elements, which were shown to corrupt the image similarity in [6]. Second, assuming the compound Poisson distribution as a good generative model of Fisher Vectors, the power transformation can be interpreted as a variance stabilizing transform, which corrects the dependence between the variance and the mean. The authors have applied the power transformation with $\alpha = 0.5$. However, the authors used only one descriptor to justify their conclusions and they did not show the impact of the power with different values of α. In the following, we will study the impact of the α parameter with the power transformation on different descriptors. Another interpretation of the power normalization benefit is that power transformation with α values smaller than 1 increases the contrast for low absolute values of the vector components. This can be seen as similar to a gamma transformation on image intensity to better show the details in dark regions of the images.

B. PCA-based dimensionality reduction

Principal Component Analysis (PCA) [7] is based on a Singular Value Decomposition (SVD) of the X matrix as:

$$X = U\Sigma V^T$$

where, U is the $d \times m$ matrix of eigenvectors of XX^T; Σ is the $m \times m$ diagonal matrix containing the square root of the eigenvalues of XX^T; V is the $n \times m$ matrix of eigenvectors of X^TX; m is the rank of XX^T and X^TX; it can also be taken as equal to d by padding the diagonal of Σ with zeros (if $d \leq n$).

In this representation, the diagonal entries in Σ are the singular values and they are normally ordered with the largest singular value (largest eigenvalue) first. Dimension reduction is achieved by dropping all but the first k of these singular values. This gives an approximation of the original matrix as:

$$X' = U'\Sigma'V'^T$$

Where U', Σ' and V' are truncated versions of U, Σ and V where only k rows and/or columns out of m were kept.

As eigenvectors are orthonormal, the Euclidean distances between the columns of X and between the same columns of $Y = XV = U\Sigma$ are identical. This is the same for the Euclidean distances between the columns of X' and between the same columns of $Y' = X'V' = U'\Sigma'$. As X' is an approximation of X, the Euclidean distances between the columns of X are approximation of the Euclidean distances between the same columns of Y'. Y' is also the same as Y in which only the k first rows are kept. The rows in the Y matrix correspond to the principal axes of the covariance matrix sorted in decreasing variance values. Dropping the last components usually does not loose much information (the last ones may be null or very small if many original components are highly correlated); it may even be beneficial since the information associated to small variance axes is generally quite noisy. In practice, it is observed that dropping the last components, an often a large proportion of them, increases the overall system performance both for indexing (classification) or retrieval (query by example or relevance feedback) [8]. There is an optimal number of components to be kept which can be determined by cross-validation within a development collection. In practice also, the mean of the column vectors of X is subtracted to all the columns before applying the PCA; subtraction the mean also does not change the Euclidean distance between column vectors.

C. Post-PCA transformations

A second normalization or transformation can also be applied after the PCA. This includes all the five previously mentioned plus another one called "whitening". L_1 and L_2 unit length can be applied to rescale the vectors and min_max and σ_norm can be applied to rescale the components. Post-PCA whitening [9] consists in dividing each component by its variance and is equivalent as such as a σ_norm. It also consist in taking $Y = XV\Sigma^{-1} = U$ from which Y' can still be obtained by dropping components. Whitening is also equivalent to replacing the Euclidean distance between the samples by the Mahalanobis distance [10]. As Mahalanobis distance is known to become noisy as low variance components are highly scaled, an improved version of the whitening can be obtained by enforcing a minimum ratio B between the first and following variances, thereby bounding the scaling of small variance components [9]. Also, some intermediate state between the Euclidean and the (improved) Mahalanobis distances can be considered, for instance by dividing by σ_i^{β} with $0 \leq \beta \leq 1$ with β controlling the "strength" of the whitening.

III. Proposed method

The proposed method for descriptor optimization is a combination of a PCA-based dimensionality reduction combined with pre-PCA and post-PCA power transformations. Though the PCA and the non-linear pre-and post-transformations have been used separately before, their specific combination and joint optimization as proposed here have not been tried before.

The main goal of the first power transformation is to transform a descriptor suited for a non-Euclidean distance (typically χ^2 for histogram / BoW-based descriptors) into a descriptor suited for an Euclidean distance with a similar or improved performance. The α_1 exponent in the first power transformation is optimized by cross-validation for obtaining the best possible performance with the Euclidean distance. The dimensionality reduction is also made by optimizing by cross-validation the k value for obtaining the best performance or the best dimensionality reduction - performance loss compromise.

For the post-PCA transformation, we propose to use a second power transformation rather than whitening. Both have the same effect of increasing small values relatively to large values, and thus preventing components with high amplitude from dominating and masking those with low amplitude, but the whitening does it based on the overall variance of a given component, while the power transformation does it on any element, regardless of it belongs to a small variance component or not. The power transformation will have effect within the values of a given small variance component, while the whitening will act globally on all of them, regardless of whether they are small or very small.

As these transformations are made sequentially, it is possible to optimize the corresponding parameters either sequentially or globally. Optimizing them jointly is more costly but it may lead to a better global performance.

IV. Experimental results

The experiments on the normalization method of video description were conducted on the TRECVid 2010 collection. This data collection consists of two large sets: the development and the test set. The development set consists of 119,685 shots of 3,173 videos with average of 37 shots per video, and the test set consists of 146,788 shots of 8,467 videos with average of 17 shots per video.

A. Video descriptors

We have used several descriptors of different types and sizes, which have been produced and shared by various partners of the IRIM project of GDR-ISIS [11]. Most of the selected descriptors are based on the color histograms or on the bag of words approaches. However, we choose to compare the methods also with different types of descriptors, such as those based on Gabor filter and audio. In practice, we have used 12 descriptors that indicates in table I. Here we detail the used descriptors:

- **global_labm1x3x192 and global_qwm1x3x192:** concatenated histogram features [12], where: "lab" refers to the use of CIE L*a*b* colors, and "qw" to the use of the quaternionic wavelets (3 scales and 3 orientations). The histogram is calculated for 3 vertical parts, and the dictionary size is 192. Both descriptors have 576 dimensions.
- **sm462:** the Saliency Moments (SM) feature [13], is a holistic descriptor that embeds locally-parsed information, namely the shape of the salient region, in a holistic representation of the scene, structurally similar to the work presented in [14]. The resulting signature vector is a 462-dimensional descriptor.
- **AudioSpectroN_b28:** Spectral profile in 28 bands on a Mel scale, N: normalized $\rightsquigarrow$ 28 dimensions.
- **dense_sift_k512:** Bag of SIFT computed with k-bin histograms, thus it has 512 dimensions.
- **h3d64:** normalized RGB Histogram $4 \times 4 \times 4 \rightsquigarrow 64$ dimensions.
- **gab40:** normalized Gabor transform, 8 orientations $\times$ 5 scales, results in 40 dimensions.
- **hg104:** early fusion of h3d64 and gab40 $\rightsquigarrow$ 104 dimensions.
- **opp_sift_<method>[_unc]_1000:** bag of visual word, opponent SIFT, generated using [15] software. <**method**> method is related to the way by which SIFT points are selected: **har** corresponds to a filtering via a Harris-Laplace detector and **dense** corresponds to a dense sampling; the versions with **_unc** correspond to the same with fuzziness introduced in the histogram computation. We have used four different descriptors of this type. The vocabulary size was of 1000.

B. Parameter optimization

For the evaluation of each normalization method, we use a multi-learner approach based SVM with RBF kernel (MSVM) as a classifier [3]. The parameters to be optimized are: the hyper-parameters if the classifier (γ) and the α parameter of the power transformations. The optimization is done by the cross-validation on the development set of TRECVid 2010, in which we split the dataset into two sets: the training and validation sets. In the following, we present the optimization process of these two parameters.

C. Evaluation of baseline normalization methods

The two vector normalization methods (L_1 and L_2) and the two component normalization methods (min_max and σ_norm) can be evaluated separately or as combined, one of each type in two possible orders.

Tables I and II show the system performance on the development set of TRECVid 2010, respectively with the Euclidean and χ^2 distance, using the four baseline normalization methods and some combinations of them (other were tried but appeared less efficient). The results obtained after normalization are compared with the result when the baseline method is used with the both distances, with no normalization at all. As we can see in these tables, the system performance varies significantly with the different normalizations. For the L_1, L_2, σ_norm and min_max normalization, the performance is in most cases closer to the baseline method, and the best normalization among them is not stable across the descriptors. The χ^2 distance is more efficient than the Euclidean one with the best baseline normalizations for the histogram-based descriptors (last eight) but does not make a significant difference for the non-histogram-based ones (first four). Other non-linear transformations, e.g. $x \leftarrow \log(1 + \alpha x)$ were also considered but none made a significant difference (not shown) probably because the difference would involve a second order effect that cannot be reliably learnt.

D. Evaluation of the power transformation

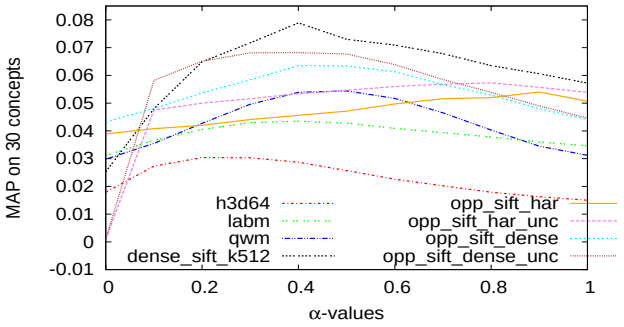
The power transformation is evaluated in conjunction with the best baseline combination for each descriptor.

TABLE I. MAP VALUES ON THE TRECVID 2010 DEVELOPMENT SET, USING THE BASELINE NORMALIZATION METHODS WITH THE EUCLIDEAN DISTANCE.

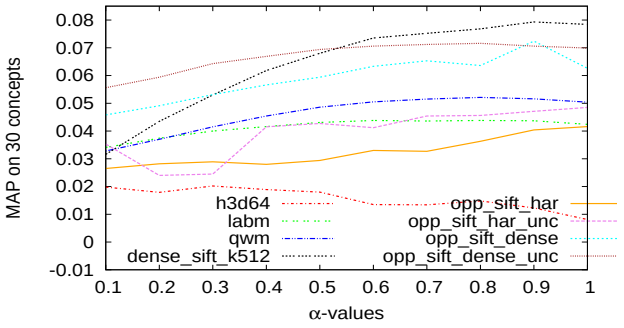
Descriptor	Raw	L_2	σ_norm	Min_max	L_2-σ_norm	σ_norm-L_2	L_2-Min_max	Min_max-L_2
sm462	0.0095	0.0121	**0.0189**	0.0115	**0.0317**	0.0235	0.0152	0.0115
AudioSpectroN_b28	0.0155	0.0156	0.0138	**0.0157**	0.0148	0.0133	**0.0158**	0.0154
vv gab40	**0.0265**	0.0257	0.0240	0.0182	**0.0267**	0.0250	0.0195	0.0142
hg104	0.0368	0.0366	**0.0407**	0.0278	0.0408	**0.0421**	0.0323	0.0284
h3d64	0.0158	0.0159	**0.0255**	0.0161	0.0227	**0.0248**	0.0152	0.0160
global_labm1x3x192	0.0346	0.0342	0.0316	**0.0355**	0.0326	0.0346	0.0348	**0.0359**
global_qwm1x3x192	0.0312	0.0351	0.0356	**0.0373**	0.0376	**0.0469**	0.0362	0.0437
dense_sift_k512	0.0572	0.0610	**0.0695**	0.0636	0.0684	**0.0733**	0.0676	0.0666
opp_sift_har_1000	0.0507	**0.0529**	0.0485	0.0455	0.0470	0.0472	0.0469	**0.0500**
opp_sift_har_unc_1000	0.0539	**0.0540**	0.0510	0.0516	0.0514	0.0504	0.0513	**0.0517**
opp_sift_dense_1000	0.0441	0.0449	**0.0545**	0.0494	0.0499	**0.0559**	0.0511	0.0507
opp_sift_dense_unc_1000	0.0446	0.0472	**0.0617**	0.0591	0.0534	**0.0626**	0.0548	0.0599

TABLE II. MAP VALUES ON THE TRECVID 2010 DEVELOPMENT SET, USING THE BASELINE NORMALIZATION METHODS WITH THE χ^2 DISTANCE.

Descriptor	Raw	L_2	σ_norm	Min_max	L_2-σ_norm	σ_norm-L_2	L_2-Min_max	Min_max-L_2
sm462	0.0144	0.0155	**0.0243**	0.0149	**0.0315**	0.0219	0.0192	0.0136
AudioSpectroN_b28	0.0030	0.0019	0.0017	**0.0096**	0.0031	0.0033	**0.0150**	0.0125
gab40	**0.0247**	0.0215	0.0240	0.0186	**0.0244**	0.0238	0.0192	0.0149
hg104	0.0378	0.0387	**0.0447**	0.0350	0.0467	**0.0477**	0.0363	0.0333
h3d64	0.0081	0.0124	**0.0137**	0.0112	**0.0326**	0.0299	0.0227	0.0254
global_labm1x3x192	0.0424	0.0399	0.0379	**0.0435**	0.0390	0.0380	**0.0423**	0.0394
global_qwm1x3x192	**0.0504**	0.0455	0.0417	0.0430	0.0415	**0.0491**	0.0439	0.0476
dense_sift_k512	0.0784	0.0760	0.0762	**0.0841**	0.0773	0.0814	**0.0820**	0.0798
opp_sift_har_1000	**0.0416**	0.0370	0.0367	0.0334	0.0460	0.0448	**0.0467**	0.0466
opp_sift_har_unc_1000	**0.0485**	0.0453	0.0425	0.0432	**0.0513**	0.0499	0.0511	0.0512
opp_sift_dense_1000	0.0623	**0.0626**	0.0586	0.0572	0.0526	**0.0563**	0.0537	0.0546
opp_sift_dense_unc_1000	0.0699	**0.0746**	0.0688	0.0676	0.0573	**0.0652**	0.0580	0.0614



(a) Euclidean distance



(b) χ^2 distance

Fig. 1. Tuning α parameter of the power transformation on the dev set of TRECVid 2010. The plots show only results with descriptors based on histogram / BoW.

Results of the α optimization are given in figure 1, for the two considered distances: Euclidean is presented in subfigure 1(a) and χ^2 in 1(b). Each curve in the plots refers to the system performance using one descriptor. As we can see, the α parameter has different values for each descriptor, with both distances. Since we believe that the only difference in the results for each descriptor and the used distance is the value of α parameter, this shows the importance of choosing the best value of α. For instance, with the Euclidean distance the h3d64 descriptor has the best performance with $\alpha = 0.3$, the dense_sift_k512 descriptor has the highest performance with $\alpha = 0.4$. Interestingly, the optimal alpha values for the χ^2 distance are approximately twice those for the Euclidean one. The optimal values for the Euclidean distance are often close to 0.5, which is a commonly used value, but not always. Detailed results per descriptor are given in the next section on the test set when parameter tuning is done by cross-validation in the development set.

E. Evaluation PCA dimensionality reduction

Figure 2 shows the system performance (MAP) obtained by applying the power transformation followed by PCA dimension reduction, with all the considered descriptors. For those with a small dimensionality, the use of the PCA is not that important. The main objective is to show the performance when using a PCA-based dimensionality reduction on high dimensionality descriptors. We have tuned the k number of PCA (i.e. number of important components) on each of the considered descriptors, using fractions from 0.1 to 1 of the original dimensions. As we can see in the figure 2, the number of the important components, varies for each of the descriptors. For long descriptors, we have fixed the k after PCA to be the value of the first fraction that has a higher performance or closer to the performance of the original dimension. For instance, the chosen k-components for the best descriptor (i.e. dense_sift_k512) is $0.4 \times 512 = 204$.

F. Evaluation of the power transformation with PCA dimensionality reduction and post-PCA transformation

While the previous results were shown for the analysis of the relevant parameter or parameter combinations within the development set, we summarize here the global results for the proposed method on the test set.

We have evaluated the different combined transformation methods on the TRECVid 2010 test set after having optimized all the relevant parameters by cross-validation for each descriptor and testing condition separately. The testing

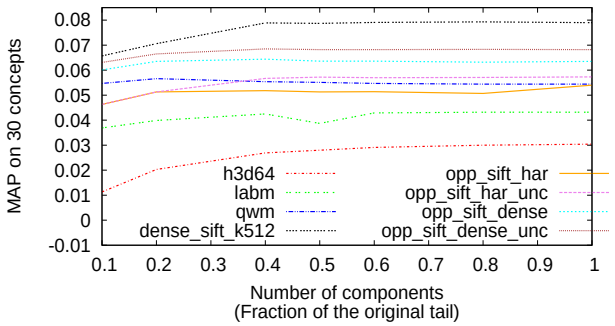


Fig. 2. Evaluating the PCA with Euclidean distance on TRECVid 2010 dev set. The plots show only results with descriptors based on histogram / BoW.

conditions included: baseline (best combination of baseline normalization); the same with pre-PCA power transformation; the previous one with PCA-based dimensionality reduction; and the previous with either post-PCA power transformation or post-PCA whitening. Results are presented for both the Euclidean and χ^2 distances for the first two and only for the Euclidean distance after PCA (since χ^2 distance in meaningless after PCA).

TABLE III. OPTIMAL VALUES OF THE PROPOSED NORMALIZATION FOR EACH DESCRIPTOR USING THE EUCLIDEAN DISTANCE.

Descriptor	d	α_1	k (σ_k^2/σ_d^2)	α_2	(B, β)
sm	462	0.200	277 (0.999)	0.6	(4.0, 0.4)
AudioSpectroN_b	28	0.200	28 (1.000)	0.8	(2.5, 0.5)
gab	40	0.300	40 (1.000)	0.6	(8.0, 0.6)
hg	104	0.300	104 (1.000)	0.6	(8.0, 0.6)
h3d	64	0.300	64 (1.000)	0.6	(4.0, 0.6)
global_labm1x3x192	576	0.400	346 (0.980)	0.5	(4.0, 0.7)
global_qwm1x3x192	576	0.500	115 (0.931)	0.7	(2.0, 0.1)
dense_sift_k	512	0.400	204 (0.931)	0.8	(2.0, 0.4)
opp_sift_har	1000	0.900	400 (0.734)	0.7	(2.0, 0.9)
opp_sift_har_unc	1000	0.800	500 (0.832)	0.8	(2.0, 0.9)
opp_sift_dense	1000	0.400	400 (0.827)	0.8	(2.0, 0.9)
opp_sift_dense_unc	1000	0.400	400 (0.933)	0.8	(2.5, 0.4)

Depending upon the testing conditions, the following hyper-parameters were optimized: α_1 the pre-PCA power transformation exponent; k the optimal number of component to keep after PCA; α_2 the post-PCA power transformation exponent; and (B, β), the optimal whitening parameters. Table III displays the optimal values found for these parameters for each considered descriptor. d is the original dimensionality of the descriptor. σ_k^2/σ_d^2 is the fraction of the kept variance in dimensionality reduction.

Results on the 12 considered descriptors are displayed in table IV. They are consistent with those on the development set (not all shown). The table shows the effectiveness of the power-law normalization (+pw) with the both distances. It further shows the effectiveness of the PCA dimensionality reduction with the Euclidean distance and the performance after PCA (+pca) using two normalizations: a second power-law and a whitening normalization (+wh). Fusion4 correspond to the performance obtained by the late fusion of the non-histogram-based descriptors scores; Fusion8 is the same with only histogram-based descriptors; Fusion-all is the same with all the considered descriptors; and Re-ranking is Fusion-all after re-ranking using the temporal context. The power transformation performs better than all the other evaluated methods for normalization for all considered descriptors. It is also better with the Euclidean distance than with the χ^2 in most cases. The use of PCA-based dimensionality reduction makes the system faster while preserving or increasing the system performance. A second power transformation improves the performance more than whitening in most cases and globally.

Results are also displayed for a system that makes a simple fusion (average of the classification scores). The fusion was tried separately for non-histogram-based and histogram-based descriptors as well as for all descriptors. The power transformation performs better with the fusion and the global one reaches the score of 0.0707 with PCA and the Euclidean distance and even 0.0807 after re-ranking using the temporal context [16]. This can be compared to the performance of the best system evaluated at TRECVid 2010 (SIN) that was 0.0900 considering that: more descriptors could have been used; the fusion method was basic; and further post-processing of the fused classification can further improve the performance, for instance using also the conceptual context [17].

G. Processing times

All the experiments were done on a machine which has two quad-core processors running at 2.66 GHz and 32 Gbytes of Ram. The execution time depends upon the size of the descriptor. It has been measured for the cumulated training and indexing times for all 30 concepts and 12 considered descriptors in hours. Using always the optimal power transformations, the total execution time is of 201 hours (using eight cores) for the optimal χ^2 distance, it is of 110 hours using the Euclidean distance, and of 53 hours with PCA dimensionality reduction. Meanwhile, as we saw before, the performance is not significantly affected or it is even improved.

H. Application to high-dimensionality descriptors

The proposed method was applied at large scale in the context of TRECVid 2012 semantic indexing task. The collection contains 545,923 video shots, 400,289 for development and 145,634 for test and 346 concepts have to be classified. IRIM produced tens of different descriptors types, many of them with variants (e.g. in the size of the BoW dictionary) resulting into over 100 descriptors, most of them being of very good quality. The dimensionality of these descriptors ranged from 15 up to 32,768. Our approach has been applied to most of them and the observed results on TRECVid 2010 data were confirmed. The optimal k was always found to be much smaller than the original size d and did not exceed 768 even for the $d \geq 10K$. This indicates that these very high-dimensional descriptors are highly redundant and have only about a few hundred useful independent components. This dimensionality reduction was again obtained with a simultaneous increase in classification performance. The overall system performance after optimized late fusion combined with the use of temporal and conceptual contexts was of 0.2692 (MAP) for our best submission while the best submitted system at TRECVid SIN 2012 was of 0.3220. The system performance was further increased to 0.3014 when more classification results were finally included.

Though not shown in this paper, additional experiments on other TRECVid and non-TRECVid video collections show that the approach is also valid with other types of video contents and other target concepts. The optimal hyper-parameter values are also quite stable across collections.

V. CONCLUSIONS

We have proposed and evaluated a method for optimizing descriptors used for content-based multimedia indexing and

TABLE IV. MAP VALUES ON THE TRECVID 2010 TEST SET, USING THE DIFFERENT NORMALIZATION METHODS WITH EUCLIDEAN OR χ^2 DISTANCES.

Descriptor	Euclidean (baseline)	χ^2 (baseline)	Euclidean +pw	χ^2 +pw	Euclidean +pw+pca	Euclidean +pw+pca+pw	Euclidean +pw+pca+wh
sm462	0.0104	0.0057	0.0246	0.0178	0.0233	**0.0339**	0.0273
AudioSpectroN_b28	0.0007	0.0007	0.0011	0.0006	0.0035	**0.0036**	0.0033
gab40	0.0106	0.0103	0.0115	0.0104	0.0114	**0.0146**	0.0144
hg104	0.0177	0.0214	0.0246	0.0207	0.0240	0.0276	**0.0285**
Fusion4	0.0295	0.0307	0.0366	0.0316	0.0403	**0.0499**	0.0461
h3d64	0.0053	0.0054	**0.0145**	0.0046	0.0126	0.0140	0.0127
global_labm1x3x192	0.0126	0.0238	0.0270	**0.0288**	0.0265	0.0275	0.0255
global_qwm1x3x192	0.0142	0.0213	0.0217	0.0227	0.0214	**0.0241**	0.0213
dense_sift_k512	0.0389	0.0420	0.0418	0.0377	0.0405	**0.0456**	0.0444
opp_sift_har_1000	0.0228	0.0187	0.0223	0.0154	0.0249	**0.0254**	0.0233
opp_sift_har_unc_1000	0.0268	0.0284	0.0293	0.0260	**0.0313**	0.0310	0.0309
opp_sift_dense_1000	0.0332	0.0340	0.0375	0.0346	0.0381	**0.0399**	0.0376
opp_sift_dense_unc_1000	0.0381	0.0433	0.0433	0.0451	0.0426	**0.0457**	0.0434
Fusion8	0.0476	0.0600	0.0624	0.0610	0.0625	**0.0669**	0.0651
Fusion-All	0.0523	0.0604	0.0632	0.0618	0.0646	**0.0707**	0.0688
Re-ranking	0.0624	0.0674	0.0723	0.0683	0.0731	**0.0807**	0.0763

retrieval. This proposed method combines a PCA-based dimensionality reduction with pre- and post-PCA non-linear transformations. The resulting transformation is globally optimized. The produced descriptors have a much lower dimensionality while performing at least as well, and often significantly better, with the Euclidean distance than the original high dimensionality descriptors with their optimal distance. They also perform better than the same descriptors optimized by classical normalization methods like L_1 or L_2 unit length or per component range (min-max) or variance normalization and simple combinations of them.

The method has been validated and evaluated for a variety of descriptors using TRECVid 2010 semantic indexing task data. It has then be applied at large scale for the TRECVid 2012 semantic indexing task on tens of descriptors of various types and with initial dimensionalities from 15 up to 32,768. The same transformation can be used also for multimedia retrieval in the context of query by example and/or relevance feedback.

ACKNOWLEDGEMENTS

This work was partly realized as part of the Quaero Program funded by OSEO, French State agency for innovation. This work was supported in part by the french project VideoSense ANR-09-CORD-026 of the ANR. Experiments presented in this paper were carried out using the Grid'5000 experimental testbed, being developed under the INRIA ALADDIN development action with support from CNRS, RENATER and several Universities as well as other funding bodies (see https://www.grid5000.fr). The authors wish to thanks the participants of the IRIM (Indexation et Recherche d'Information Multimédia) group of the GDR-ISIS research network from CNRS for providing the descriptors used in these experiments.

REFERENCES

[1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2*, ser. ICCV '03. Washington, DC, USA: IEEE Computer Society, 2003, pp. 1470–. [Online]. Available: http://dl.acm.org/citation.cfm?id=946247.946751

[2] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1–22.

[3] B. Safadi and G. Quénot, "Active learning with multiple classifiers for multimedia indexing," *Multimedia Tools and Applications*, September 2010.

[4] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[5] H. Jégou, F. Perronnin, M. Douze, J. Sánchez, P. Pérez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011.

[6] H. Jégou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR '09)*. Miami, United States: IEEE Computer society, 2009, pp. 1169–1176. [Online]. Available: http://hal.inria.fr/inria-00394211

[7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, 1st ed. Springer, October 2007.

[8] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *23rd IEEE Conference on Computer Vision & Pattern Recognition (CVPR '10)*. San Francisco, United States: IEEE Computer Society, 2010, pp. 3304–3311.

[9] H. Jégou and O. Chum, "Negative evidences and co-occurrences in image retrieval: the benefit of PCA and whitening," in *ECCV - European Conference on Computer Vision*, Firenze, Italie, Oct. 2012.

[10] P. C. Mahalanobis, "On the generalised distance in statistics," *Proceedings National Institute of Science, India*, vol. 2, no. 1, pp. 49–55, Apr. 1936.

[11] D. Gorisse, F. Precioso, P. Gosselin, L. Granjon, D. Pellerin, M. Rombaut, H. Bredin, L. Koenig, H. Lachambre, E. El Khoury, R. Vieux, B. Mansencal, Y. Zhou, J. Benois-Pineau, H. Jégou, S. Ayache, B. Safadi, Y. Tong, F. Thollard, G. Quénot, A. Benoit, and P. Lambert, "IRIM at TRECVID 2010: High Level Feature Extraction and Instance Search," in *TREC Video Retrieval Evaluation workshop*. Gaithersburg, MD USA: National Institute of Standards and Technology, nov 2010.

[12] D. Gorisse, M. Cord, and F. Precioso, "Salsas: Sub-linear active learning strategy with approximate k-nn search," *Pattern Recognition*, vol. 44, no. 10-11, pp. 2343–2357, 2011.

[13] M. Redi and B. Merialdo, "Saliency moments for image categorization," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ser. ICMR '11. New York, NY, USA: ACM, 2011, pp. 39:1–39:8. [Online]. Available: http://doi.acm.org/10.1145/1991996.1992035

[14] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, pp. 145–175, 2001.

[15] K. van de Sande, T. Gevers, and C. Snoek, "Evaluation of color descriptors for object and scene recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, june 2008, pp. 1 –8.

[16] B. Safadi and G. Quénot, "Re-ranking by local re-scoring for video indexing and retrieval," in *Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM)*, Glasgow, United Kingdom, 2011, pp. 2081–2084.

[17] A. Hamadi, G. Qunot, and P. Mulhem, "Two-layers re-ranking approach based on contextual information for visual concepts detection in videos," in *CBMI*, jun 2012, pp. 1–6.