MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Dimensionality Reduction of Visual Features for Efficient Retrieval and Classification

Boufounos, P.T.; Mansour, H.; Rane, S.D.; Vetro, A.

TR2016-103 July 2016

Abstract

Visual retrieval and classification are of growing importance for a number of applications, including surveillance, automotive, as well as web and mobile search. To facilitate these processes, features are often computed from images to extract discriminative aspects of the scene, such as structure, texture or color information. Ideally, these features would be robust to changes in perspective, illumination, and other transformations. This paper examines two approaches that employ dimensionality reduction for fast and accurate matching of visual features while also being bandwidth-efficient, scalable and parallelizable. We focus on two classes of techniques to illustrate the benefits of dimensionality reduction in the context of various industrial applications. The first method is referred to as quantized embeddings, which generates a distance-preserving feature vector with low rate. The second method is a low rank matrix factorization applied to a sequence of visual features, which exploits the temporal redundancy among features vectors associated with each frame in a video. Both methods discussed in this paper are also universal in that they do not require prior assumptions about the statistical properties of the signals in the database or the query. Furthermore, they enable the system designer to navigate a rate vs. performance trade-off similar to the rate-distortion trade-off in conventional compression.

APSIPA Transactions on Signal and Information Processing

Dimensionality Reduction of Visual Features for Efficient Retrieval and Classification

Petros T. Boufounos, Hassan Mansour, Shantanu Rane and Anthony Vetro

Abstract

Visual retrieval and classification are of growing importance for a number of applications, including surveillance, automotive, as well as web and mobile search. To facilitate these processes, features are often computed from images to extract discriminative aspects of the scene, such as structure, texture or color information. Ideally, these features would be robust to changes in perspective, illumination, and other transformations. This paper examines two approaches that employ dimensionality reduction for fast and accurate matching of visual features while also being bandwidth-efficient, scalable and parallelizable. We focus on two classes of techniques to illustrate the benefits of dimensionality reduction in the context of various industrial applications. The first method is referred to as quantized embeddings, which generates a distance-preserving feature vector with low rate. The second method is a low rank matrix factorization applied to a sequence of visual features, which exploits the temporal redundancy among features vectors associated with each frame in a video. Both methods discussed in this paper are also universal in that they do not require prior assumptions about the statistical properties of the signals in the database or the query. Furthermore, they enable the system designer to navigate a rate vs. performance trade-off similar to the rate-distortion trade-off in conventional compression.

Index Terms

Randomized Embeddings, Nearest Neighbors, Quantization, Low Rank Matrix Factorization, Visual Search, Classification

P. T. Boufounos, H. Mansour and A. Vetro are with Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA 02139. {petrosb,mansour,avetro}@merl.com.

S. Rane is with Palo Alto Research Center (PARC), Palo Alto, CA 94304. srane@parc.com. S. Rane was with Mitsubishi Electric Research Laboratories (MERL) when this work was performed.

Dimensionality Reduction of Visual Features for Efficient Retrieval and Classification

I. Introduction

The amount of visual data generated by humans continues to grow at a staggering pace. This has created a plethora of new applications that were not conceivable even a decade ago: face tagging in images uploaded to social networking sites, extraction of rich information about products photographed at a supermarket, geographical and historical data mining about landmarks photographed on a touristic excursion, and augmented reality, to name a few. The increased diversity and redundancy of today's rapidly growing databases enables the development of robust and novel applications with unprecedented capabilities.

Visual search and retrieval has been an active area of research for decades and there is a large body of work on image descriptors that enable fast and accurate image retrieval. Some popular descriptors include SIFT [1], SURF [2], GIST [3], BRISK [4] and FREAK [5]. Of these, GIST captures global properties of the image, while the others capture local details at several salient points in an image, and therefore, have been used to match local features or patches. These descriptors can be used for image matching, registration and retrieval by combining hypotheses from several image patches, for example, using the popular Bag-of-Features approach [6]. A comparative study of image descriptors has shown that Scale Invariant Feature Transformation (SIFT) features have the highest robustness against common image deformations such as translation, rotation, and a limited amount of scaling [7]. However, recent work has reported FREAK to outperform SIFT in terms of robustness and speed [5].

There are a variety of other features that are well suited for specific visual inference tasks. For example, eigenfaces [8] or Viola-Jones face descriptors [9] are typically preferred for face recognition, while the matching of other biometrics, such as fingerprints, requires features specifically designed to maximize matching performance for a given biometric sensor [10]. In recent years, deep learning has also become a popular means for determining the best features for a given inference task, e.g., [11].

For many problems, the sheer size of the data to be searched makes image-based querying extremely challenging, especially in bandwidth-restricted applications that depend upon the speed of information retrieval. This problem is further compounded by the size of the search query. For example, a SIFT feature vector for a single salient point in an image is a real-valued, unit-norm 128-dimensional vector.

Finding a reliable match between a server-side image and a photograph sent by a mobile client typically requires features obtained from several hundred salient points. Therefore, a prohibitively large bit rate is required to transmit the SIFT features from a client device to a database server for the purpose of matching. So, while feature selection is a key issue for many visual inference tasks, the focus of this paper is on the efficiency of the storage, transmission and matching processes.

It goes without saying that the visual inference mechanism must identify visual matches accurately. However, the accuracy requirement cannot be considered in isolation, especially when tradeoffs need to be made to ensure practical feasibility of the algorithm. Driven by various industrial applications, we believe it is important that the following requirements also be satisfied:

1) *Compact upload from the client device to the server or cloud*: To minimize the communication overhead, the visual inference mechanism must ensure that the client sends the query signal to the server using the smallest possible number of bits. Small query vectors will help to satisfy the low latency requirements for real-time applications, as further elaborated below.
2) *Fast search algorithm at the server*: In several application scenarios, the information retrieved by the visual inference mechanism may be time-sensitive. Examples include augmented reality-enabled headsets for people browsing museum exhibits, or scene-specific route guidance for cars, etc. In such cases, it is beneficial for the server-based matching algorithm to be as fast and parallelizable as possible without compromising the matching performance.
3) *Robustness to variations in the visual query*: For most interesting applications of visual inference, there is no guarantee that the images taken by the client device will be optimally aligned with the images in the server's database. For example, the server's database may be compiled via crowdsourcing from a very large number of users. Therefore, it is imperative that the visual search be robust to variations in the input image, such as, changes in translation, rotation, scaling, image quality, illumination, and so on.
4) *Futureproof algorithm*: In many visual inference applications, the server's database keeps changing as new visual data is added and low-quality or irrelevant data is discarded. For instance, the performance of a mobile phone-based augmented reality application would improve as a richer variety of images are accumulated at the database server with time. From a practical perspective, it is desirable that the algorithm and parameters used by the client remain unchanged when the server's database changes. Frequently changing the parameter values may well guarantee optimal performance, but would also require the client device to download frequent updates containing the

retrained parameters.

This paper considers the utilization of dimensionality reduction techniques to address the above issues in the context of several retrieval and classification tasks. Two distinct approaches are described to illustrate the benefits. The first method is referred to as quantized embeddings and generates randomized distance-preserving embeddings of image features. Specifically, we use a randomized linear transformation, known as a Johnson-Lindenstrauss embedding [12], to map the image features to a lower-dimensional space, followed by coarse, possibly dithered, quantization. The embedded signals are transmitted to the server to perform inference or matching. We show that the reduced dimensionality, combined with the coarse quantization, can significantly reduce the necessary bandwidth while maintaining matching performance. The second method is a low rank matrix factorization applied to a sequence of visual features, which exploits the temporal redundancy among feature vectors associated with each frame in a video. As with the embeddings, we demonstrate that the rate of the visual features can be dramatically reduced while maintaining classification accuracy. Both methods discussed in this paper are also universal in that there are no assumptions about the statistical properties of the signals in the database or the query. It should also be noted that while we focus on visual inference and image retrieval, the described methods are also applicable to a wide variety of other modalities and applications.

The remainder of the paper is organized as follows. Section II analyzes the performance of embeddings under quantization and demonstrates how quantized embeddings can be used to convert scale-invariant features into low bit-rate representations that are suitable for visual search applications. Section III discusses the use of low rank matrix factorization techniques to extract compact features from a large sequence of (possibly dense) visual features. Additional methods and related work are discussed in Section IV. Section V provides some concluding remarks.

II. Quantized Embeddings of Feature Spaces

We are interested in an efficient, general method for inference problems including search, retrieval and classification of visual data, such as images and videos. For the search to be accurate, it is necessary to choose an appropriate feature space with a high matching accuracy. Having chosen the feature space, a dimensionality reduction step typically follows, resulting in a descriptor. This is a critical step in reducing both the bandwidth and the search time of the query. This section describes several design approaches, collectively referred to as quantized embeddings, which reduce the bit rate of the feature vectors while maintaining the performance of the inference task.

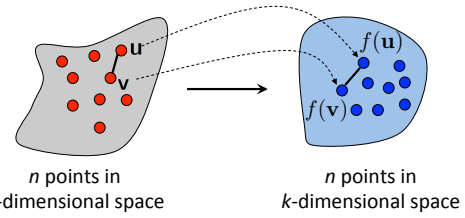


Fig. 1. The Johnson-Lindenstrauss Lemma guarantees the existence of an embedding that preserves pairwise Euclidean distances.

A. *Randomized Embeddings*

An embedding is a mapping of a set $\mathcal{X}$ to another set $\mathcal{V}$ that preserves some property of $\mathcal{X}$ in $\mathcal{V}$. Embeddings enable algorithms to operate on the embedded data, allowing processing and inference, so long as the processing relies on the preserved property. In the context of visual retrieval and classification, we consider applying a randomized embedding of the features extracted from the image or video of interest in order to reduce the dimensionality of the feature vector.

The use of embeddings is justified by the Johnson-Lindenstrauss (JL) lemma, which states that one can design an embedding $f(\cdot)$ such that for all pairs of signals $\mathbf{u}, \mathbf{v} \in \mathcal{X} \subset \mathbb{R}^d$, their embedding, $f(\mathbf{u}), f(\mathbf{v}) \in \mathbb{R}^k$ satisfies:

$$(1-\epsilon)\|\mathbf{u}-\mathbf{v}\|_2^2 \leq \|f(\mathbf{u})-f(\mathbf{v})\|_2^2 \leq (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|_2^2 \tag{1}$$

for some ϵ, as long as $k = O\left(\frac{\log P}{\epsilon^2}\right)$, where P is the number of points in $\mathcal{X}$.

A key feature of the JL lemma is that the embedding dimension k depends logarithmically only on the number of points in the set, and not on its ambient dimension d. It establishes a dimensionality reduction result, in which any set of n points in a d-dimensional Euclidean space can be embedded into a k-dimensional Euclidean space, as shown in Fig 1. Thus, the embedding dimension can typically be much lower than the ambient dimension, with minimal compromise on the embedding fidelity, as measured by ϵ. Any processing based on distances between signals—which includes the majority of inference methods—can thus operate on the much lower-dimensional space.

One way to construct the embedding function f is to project the points from $\mathcal{X}$ onto a random hyperplane passing through the origin, drawn from a rotationally invariant distribution. In practice, this is accomplished by multiplying the data vector with a matrix whose entries are drawn from a specified distribution. Concretely, the JL map can be realized using a linear map $f(\mathbf{u}) = \frac{1}{\sqrt{k}}\mathbf{A}\mathbf{u}$, where the $k \times d$ matrix $\mathbf{A}$ can be generated using a variety of random constructions [13], [14]. For example, it has been

shown that a matrix with i.i.d. $\mathcal{N}(0,1)$ entries provides the distance-preserving properties in Eqn. (1) with high probability.

In general, the embedding dimension depends on the complexity of the signal set. For discrete points it only depends logarithmically to their number. However, as the set becomes denser, other measures of complexity, such as the set covering number, can be used to better characterize the embedding dimension. For examples, see [15]–[17] and references within. Thus, embedding dimensionality can be kept low, even for very large, and increasing in size, datasets. Of course, with such large signal sets, it is necessary that the features are sufficiently discriminative to perform the required tasks.

Our focus in this paper is bit rate and communication complexity. However, it should also be noted that embeddings reduce complexity due to the dimensionality of the problem, not complexity due to the size of the signal set. This is often an issue with search-based techniques, such as nearest neighbors. Techniques including locality-sensitive hashing (LSH) and tree-based searches can be used to reduce query complexity in such cases, as discussed in Sec. IV. Complexity due to data size is less of an issue for trained-based techniques, such as neural networks and support vector machines (SVM). In those, the data-intensive training stage is typically performed off-line and can afford significantly more computational and storage complexity.

B. Quantized Embeddings

We now consider the problem of transforming the selected features into a compact descriptor that occupies a significantly smaller number of bits, while preserving the matching accuracy of the native feature space. We first perform the dimensionality reduction using a JL embedding, as described in the previous subsection. However, even though the dimensionality of the embedding $f(\mathbf{u})$ is smaller than the original feature vector $\mathbf{u}$, the elements of $f(\mathbf{u})$ are real-valued and thus cannot be represented using a finite number of bits. In order to make it feasible to store and transmit the distance-preserving embeddings $f(\mathbf{u})$, the real-valued random projections have to be quantized.

Consider a finite-rate uniform scalar quantizer $q(\cdot)$ as shown in Fig. 2 with stepsize Δ and S as the saturation level of the quantizer. Using such a quantizer, it was shown in [18] that a JL map $f(\mathbf{u}) = \frac{1}{\sqrt{k}}\mathbf{A}\mathbf{u}$ can be quantized to $g(\mathbf{u}) = \frac{1}{\sqrt{k}}q(\mathbf{A}\mathbf{u})$ and satisfy the following condition:

$$(1-\epsilon)\|\mathbf{u}-\mathbf{v}\|_2 - \Delta \leq \|g(\mathbf{u}) - g(\mathbf{v})\|_2 \leq (1+\epsilon)\|\mathbf{u}-\mathbf{v}\|_2 + \Delta \tag{2}$$

The above condition indicates that quantized embeddings preserve pairwise Euclidean distances up to a multiplicative factor $1 \pm \epsilon$, and an additive factor $\pm\Delta$. It follows from the JL lemma, that reducing ϵ

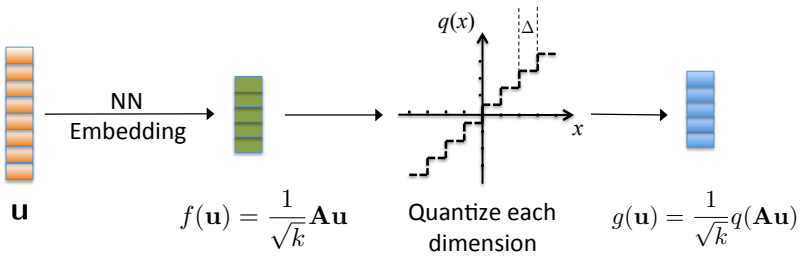


Fig. 2. (a) A quantized embedding is derived by first obtaining a Johnson-Lindenstrauss embedding by multiplying the vectors in the canonical feature space by a random matrix, followed by scalar quantization of each element in the vector of randomized measurements.

requires a greater number of randomized measurements k. Furthermore, reducing Δ amounts to increasing the number of bits allocated to each measurement. This suggests a trade-off between the embedding space dimension k, i.e., the number of measurements, and the number of bits per measurement B. For a fixed bit rate, fewer measurements and more bits per measurements will increase the error due to the JL embedding, ϵ, while a greater number of measurements and fewer bits per measurement will increase the error due to quantization, Δ. The design choice should balance the two errors. A tighter guarantee which, however, exhibits the same trade-off has also been established when the ℓ_1 distance is used in the embedding space [19].

The quantization interval Δ can be more explicitly expressed in terms of the parameters that characterize the scalar quantizer and the bits used to encode each measurement. For the finite uniform scalar quantizer shown in Fig. 2 with saturation levels $\pm S$, the quantization interval is given by $\Delta = 2^{-B+1}S$. Using R to denote the total rate available to transmit the k measurements, i.e., setting $B = R/k$ bits per measurement, the quantization interval then becomes $\Delta = 2^{-R/k+1}S$. Fig. 3 illustrates qualitatively the tradeoff between the measurement error and quantization error. Unfortunately, due to the dependence of the error on the distance between the signals, and to the existence of several loosely determined constants in the proofs of embedding theorems, this tradeoff can only be explored using experimental data. Still, we should note that ϵ scales approximately proportionally to $1/\sqrt{k}$ when small.

Thus far, we have only discussed quantized embeddings under a Euclidean (ℓ_2) distance criterion. However, extensions to non-Euclidean distances, such as the ℓ_1 distance, are also possible. For example, Indyk *et al.* [20] have described an embedding into a normed ℓ_1 metric space that preserves ℓ_1 distance, such that the distance in the embedding space within a $1-\epsilon$ factor of the original distance with *high* probability, and within a $1+\epsilon$ factor of the original distance with *constant* probability.

If we assume *integer feature vectors* with *bounded elements*, it becomes possible to preserve ℓ_1 distance

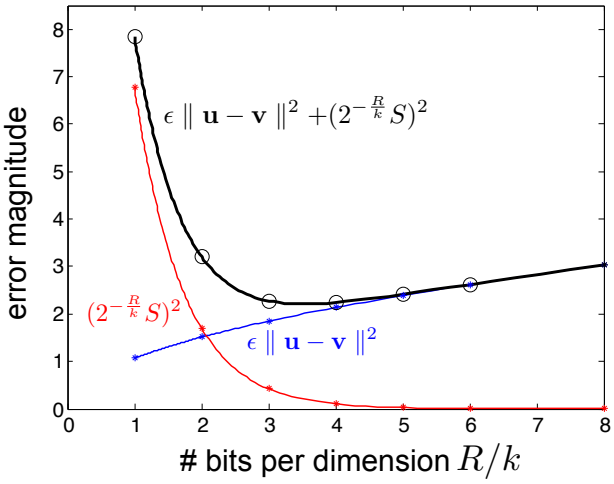


Fig. 3. Reducing the bits per dimension increases the quantization error (red curve), but allows more dimensions, thereby reducing the embedding error (blue curve). The total error plot (black curve) suggests an optimal tradeoff between the number of dimensions and the number of bits allocated per dimension.

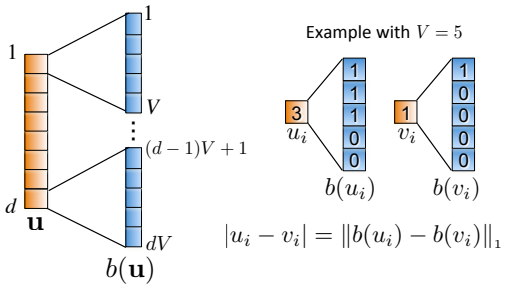


Fig. 4. "Unary" expansion of an integer vector to preserve ℓ_1 distances. Each element of the original vector is expanded to V bits, such that if the coefficient value is u_i, the first u_i bits are set to 1 and the next $V - u_i$ bits are set to zero. The ℓ_2 distance between expanded vectors equals the ℓ_1 distance between the original vectors. This requires that u_i is bounded by V. Thus a d dimensional vector is expanded to dV dimensions.

to within a $1 \pm \epsilon$ factor with high probability. This can be achieved by naively mapping the integer feature vectors into binary feature vectors, as shown in Fig. 4, using what is sometimes referred to as a "unary" expansion. The expanded vectors are elements of a real-valued ℓ_2 metric space, such that the squared ℓ_2 distance between the binary vectors is exactly equal to the ℓ_1 distance between the original integer feature vectors [18], [21].

Such a mapping, which is an isometry into a Hamming space, was first suggested in [22]. With this mapping, it then becomes possible to apply the quantized embeddings to the binary feature vectors, which has the effect of preserving the pairwise ℓ_1 distance between the original integer feature vectors. Since the embedding dimension only depends on the number of signals in the original space, the significant

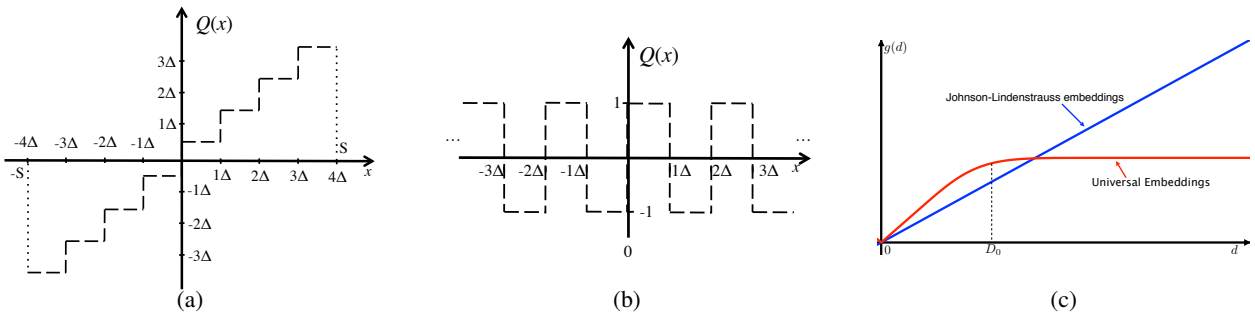


Fig. 5. (a) Conventional 3-bit (8 levels) scalar quantizer with saturation level $S = 4\Delta$. (b) Universal scalar quantizer. (c) The embedding map $g(d)$ for JL-based embeddings (blue) and for universal embeddings (red).

intermediate dimensionality expansion does not affect the embedding dimension. Results of this approach on a face verification experiment in which the underlying feature space consists of Viola-Jones face features were presented in [23].

C. Universal Embeddings

More sophisticated quantizer designs are possible within this framework, and have been shown to provide interesting tradeoffs among matching accuracy, bit rate efficiency and privacy [24]–[26]. Rather than using a finite-range uniform quantizer, an alternative approach uses a non-monotonic quantizer combined with dither, which can preserve distances up to a certain radius, as determined by the embedding parameters. Furthermore, given a fixed total rate, R, the quality of the embedding depends on the range of distances it is designed to preserve. At a fixed bit-rate, increasing the range of preserved distances also increases the ambiguity of how well the distances are preserved. Specifically, universal embeddings use a map of the form:

$$\mathbf{q} = Q(\mathbf{A}\mathbf{u} + \mathbf{w}), \tag{3}$$

where $\mathbf{A} \in \mathbb{R}^{k \times d}$ is a matrix with entries drawn from an i.i.d. standard normal distribution, $Q(\cdot)$ is the quantizer, and $\mathbf{w} \in \mathbb{R}^k$ is a dither vector with entries drawn from a $[0, \Delta]$ uniform i.i.d. distribution. An important difference with respect to conventional embeddings is that the quantizer $Q(\cdot)$ is not a conventional quantizer shown in Fig. 5(a). Instead, the non-monotonic 1-bit quantizer in Fig. 5(b) is used. This means that values that are very different could quantize to the same level. However, for local distances that lie within a small radius of each value, the quantizer behaves as a regular quantizer with dither and stepsize Δ.

In particular, universal embeddings have been shown to satisfy:

$$g\left(\|\mathbf{u}-\mathbf{v}\|_2\right)-\tau \leq d_H\left(f(\mathbf{u}), f(\mathbf{v})\right) \leq g\left(\|\mathbf{u}-\mathbf{v}\|_2\right)+\tau, \tag{4}$$

where $d_H(\cdot,\cdot)$ is the Hamming distance of the embedded signals and $g(d)$ is the map:

$$g(d)=\frac{1}{2}-\sum_{i=0}^{+\infty}\frac{e^{-\left(\frac{\pi(2i+1)d}{\sqrt{2}\Delta}\right)^2}}{\left(\pi(i+1/2)\right)^2}. \tag{5}$$

Similarly to JL embeddings, universal embeddings hold with overwhelming probability as long as $M = O\left(\frac{\log P}{\tau^2}\right)$, where, again, P is the number of points in $\mathcal{X}$.

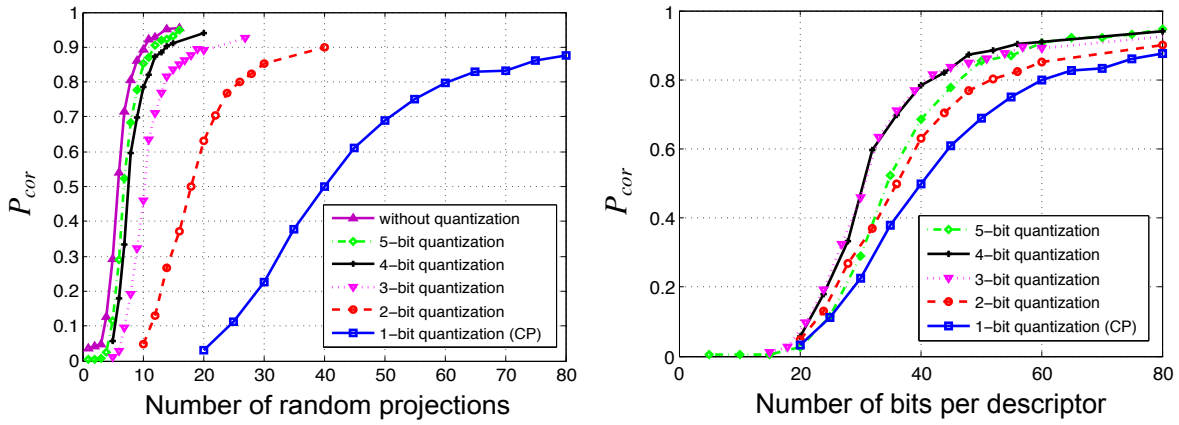
The behavior of universal embeddings is illustrated in Fig. 5(c). In particular, $g(\cdot)$ can be very well approximated by a linear portion that reaches a saturation point at distance D_0 and then saturates to a constant portion. The slope of the linear portion is determined only by the choice of Δ, which also determines the distance D_0 at which distance preservation saturates. Specifically, D_0 is proportional to Δ; a larger choice of Δ implies that a larger range of distances is preserved. On the other hand, as described in [15], [25], given a fixed rate, preserving a larger range of distances by selecting a larger Δ reduces the fidelity with which these distances are preserved.

D. Experimental Results

In this section, we discuss the performance of quantized embeddings for two application scenarios. The first application is visual inference on natural images in which we find similar images based on embeddings of SIFT features. The second application is object classification based on HOG features and SVM classification.

1) Embeddings of Scale-Invariant Features: We conducted experiments on a public database to evaluate the performance of meta-data retrieval using quantized embeddings of scale-invariant features. We used the ZuBuD database [27], which contains 1005 images of 201 buildings in the city of Zurich. There are 5 images of each building taken from different viewpoints. The images were all of size 640×480 pixels, compressed in PNG format. One out of the 5 viewpoints of each building was randomly selected as the query image, forming a query image set of $s = 201$ images. The server's database then contains the remaining 4 images of each building, for a total of $t = 804$ images.

SIFT features are extracted from the query and database images and features are matched using their quantized embeddings. The exact details, including the matching algorithm and extensive results, have been described in recent papers [15], [18], [25]. Here, we summarize the experiments that examine the



(a) (b)

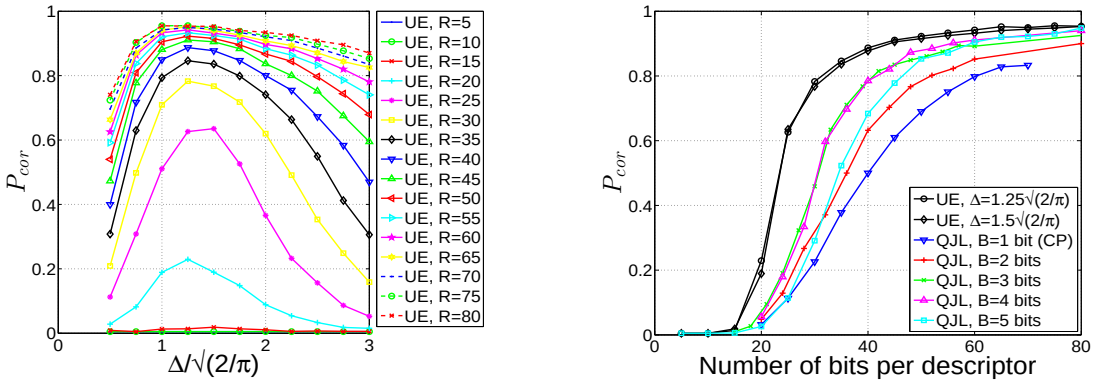
Fig. 6. (a) Multi-bit quantization with fewer random projections outperforms LSH-based schemes [28], [29] which employ 1-bit quantization with a large number of random projections. (b) When the bit budget allocated to each descriptor (vector) is fixed, the best retrieval performance is achieved with 3-bit and 4-bit quantization.

performance of our approach and the trade-off between number of measurements and number of bits per measurement with respect to that performance.

To measure the fidelity of the algorithm, we define the probability of correct retrieval P_{cor} simply as the expected value of the ratio of the number of query images for which the Nearest Neighbor (NN) search yields the correct match (N_c), to the total number of query images (N_q), which is 201 for the ZuBuD database. In this definition, the expectation is taken over the randomness in the experiment, namely the realization of the random projection matrix $\mathbf{A}$. We repeated each experiment 30 times, using a different random realization of $\mathbf{A}$ each time, reporting the mean of the ratio N_c/N_q as P_{cor}.

We first compared the accuracy of meta-data retrieval achieved by the LSH-based 1-bit quantization schemes [28], [29] with our multi-bit quantization approach. Both the LSH-based schemes use random projections of the SIFT vectors followed by 1-bit quantization according to the sign of the random projections. Fig. 6(a) shows the variation of P_{cor} against the number of projections for the LSH-based schemes. This is significantly outperformed by meta-data retrieval based on unquantized projections. Between the two extremes lie the performance curves of the multibit quantization schemes. Using 4 or 5 bits per dimension nearly achieves the performance of unquantized random projections. For the same number of measurements, this comes at a significant rate increase, compared to 1-bit measurements.

Next, we examine experimentally the optimal trade-off between number of measurements, k, and bits per measurement, B, to achieve highest rate of correct retrieval, P_{cor}, given a fixed total rate budget, $R = kB$, per embedded descriptor. This is shown in Fig. 6(b). A multibit quantizer again gives higher



(a) (b)

Fig. 7. (a) Universal embedding performance at different bit rates, as a function of Δ (b) Performance of properly tuned universal embeddings (UE) as a function of the rate, compared with the conventionally quantized JL embeddings (QJL) also shown in Fig. 6(b).

probability of correct retrieval than the 1-bit quantization schemes, confirming that taking few finely quantized projections can outperform taking many coarsely quantized projections. However, more bits per measurement are not always better. In particular, the 3 and 4-bit quantizers provide the highest P_{cor} for a given bit budget, outperforming the 5-bit quantizer.

The trade-off can be further improved using universal embeddings, as shown in Fig. 7. Specifically, Fig. 7(a) demonstrates the effect of varying the locality parameter Δ in the experiments, at various bit rates. As the radius of distances preserved expands, performance improves, as expected from the theory. However, beyond a certain radius, further expansion is not necessary for nearest neighbor identification; expanding the radius only reduces the fidelity of the embedding, thus reducing the retrieval accuracy. Figure 7(b) demonstrates the improved performance of properly tuned universal embeddings, compared to conventionally quantized JL embeddings. We should also note that tuning universal embeddings can be done in a principled way, by designing Δ according to the distances that should be preserved in the data set. In contrast, there is no principled method to design quantized embeddings, i.e., select B and k given the desired rate R.

These experiments confirm that using quantized embeddings is significantly more efficient than sending quantized versions of the original descriptor. In our experiments, the performance of quantized SIFT vectors saturated at 94%, using 384 bits per descriptor. The same performance is achieved consuming only 80 bits per descriptor using quantized embeddings of the descriptor, and 60 bits using universal embeddings. In other words, quantized and universal embeddings provide approximately 79% and 84%

lower rate compared to SIFT features, respectively, with the same retrieval performance.

Furthermore, the scheme is much more efficient than compressing the original image via JPEG and transmitting it to the server for SIFT-based matching. At 80 quality factor, the average size of a JPEG-compressed image from the ZuBuD database is 58.5 KB. In comparison, the average total bit rate of all embeddings computed for an image in this database is 2.5 KB using quantized embeddings and 1.9KB using universal embeddings.

2) Embeddings of HOG Features for SVM Classification: Next, we demonstrate the performance of compressed features on a multiclass classification problem. The goal is to identify the class membership of query images belonging to one of 8 different classes.

To set up this problem, we extract Histogram of Oriented Gradients (HOG) features [30] from 15 training and 15 test images. The HOG algorithm extracts a 36 element feature vector (descriptor) for every 8×8 pixel block in an image. The descriptors encode local histograms of gradient directions in small spatial regions in an image. Every HOG feature is compressed using either quantized JL embeddings or universal quantized embeddings. The compressed features are then stacked to produce a single compressed feature vector for each image. Then, the compressed features of the training images are used to train a binary linear SVM classifier. In the testing stage, compressed HOG features of the test images, i.e., candidate query images, are computed and classification is performed using the trained SVM classifier. In our simulations, we used tools from the VLFeat library [31] to extract HOG features and train the SVM classifier.

We consider eight image classes. One is the persons from the INRIA person dataset [30], [32]. The other seven—car, wheelchair, stop sign, ball, tree, motorcycle, and face—are extracted from the Caltech 101 dataset [33], [34]. All images are standardized to 128×128 pixels centered around the target object in each class.

Fig. 8(a) shows the classification accuracy obtained by quantized JL embeddings of HOG descriptors using the trained SVM classifier. The black square corresponds to 1-bit scalar quantization of raw non-embedded HOG descriptors, each consuming 36 bits—one bit for each element of the descriptor.

The figure shows that 1-bit quantized JL embeddings achieves a 50% bit-rate reduction, compared to non-embedded quantized descriptors, without impacting classification accuracy. This is obtained using an 18-dimensional JL embedding of every HOG descriptor, followed by 1-bit scalar quantization. Furthermore, increasing the embedding dimension, and, therefore, the bit-rate, above 18 improves the inference performance beyond that of the 1-bit quantized non-embedded HOG features. Note that, among all quantized JL embeddings, 1-bit quantization achieves the best rate-inference performance.

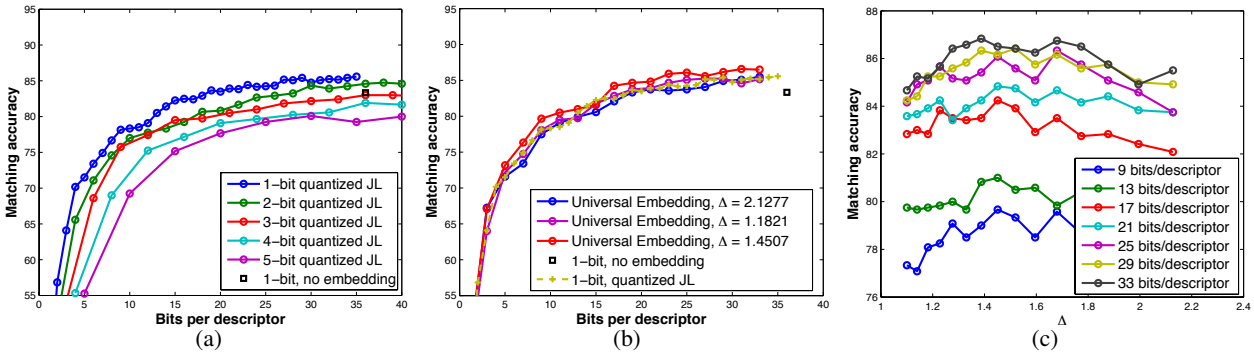


Fig. 8. Classification accuracy as a function of the bit-rate achieved using (a) quantized JL (QJL) embeddings; and (b) the universal embeddings. (c) Classification accuracy as a function of the quantization step size Δ used in computing the universal embeddings.

Fig. 8(b) compares the classification accuracy of universal embeddings for varying values of the step size parameter Δ with that of the 1-bit quantized JL embeddings and the 1-bit quantized non-embedded HOG descriptors. With the choice of $\Delta = 1.4507$, the universal embedded descriptors further improve the rate-inference performance over the quantized JL embeddings. In particular, they also achieve the same classification accuracy as any choice of quantization for non-embedded HOG descriptors, or even, unquantized ones, at significantly lower bit-rate. The datapoints representing the bit-rate vs. accuracy tradeoff for unquantized HOG descriptors are not shown in the figure, as they are out of the interesting part of the bit-rate scale.

Figure 8(c) illustrates the effect of the parameter Δ by plotting the classification accuracy as a function of Δ for different embedding rates. The figure shows that, similar to the findings in [35], if Δ is too small or too large, the performance suffers.

As evident, an embedding-based system design can be tuned to operate at any point on the rate vs. classification performance frontier, which is not possible just by quantizing the raw features. Furthermore, with the appropriate choice of Δ, universal embeddings improve the classification accuracy given the fixed bit-rate, compared with quantized JL embeddings, or reduce the bit-rate required to deliver a certain inference performance.

III. Low-Rank Matrix Factorization of Visual Features

Compact descriptors of visual scenes allow us to reduce the amount of metadata that is compressed and stored with the video bitstream while maintaining a discriminative representation of the scene content. We assume that local scene descriptors, such as SIFT or HoG features, are extracted from every video

frame in a group of pictures (GOP). The descriptors are then stacked together to form a matrix X of size $m \times N$, where m is the length of the feature vector and N is the total number of descriptors extracted from the GOP. In many situations, the number of descriptors N can reach several hundred features per frame. Therefore, it is imperative that these descriptors be encoded in a compact manner.

This section considers the problem of extracting descriptors that represent visually salient portions of a video sequence. Specifically, we describe a feature-agnostic approach for efficient retrieval of similar video content in which the extraction of compact video descriptors is cast as either a k-means clustering problem or a Non-negative Matrix Factorization (NMF) problem.

A. Background

Matrix factorization is an effective technique commonly used for finding low dimensional representations for high dimensional data. An $m \times N$ matrix X is factored into two components L, R such that their product closely approximates the original matrix

$$X \approx LR. \tag{6}$$

In the special case where the matrix and its factors have non-negative entries, the problem is known as non-negative matrix factorization (NMF). First introduced by Paatero and Tapper [36], NMF has gained popularity in machine learning and data mining following the work of Lee and Seung [37]. Several NMF formulations exist, with variations on the approximation cost function, the structure imposed on the non-negative factors, applications, and the computational methods to achieve the factorization, among others [38].

Here, we examine NMF formulations proposed for clustering [39], [40]. Specifically, we consider the sparse and orthogonal NMF formulations. The orthogonal NMF problem is defined as

$$\min_{L \geq 0, R \geq 0} \frac{1}{2}\|X - LR\|_F^2 \text{ s.t. } RR^T = I, \tag{7}$$

which was shown in [39] to be equivalent to k-means clustering. Alternatively, the sparse NMF problem [40] relaxes the orthogonality constraint on R replacing it with an ℓ_1 norm regularizer on the columns of R and a smoothing Frobenius norm on L. The sparse NMF problem is explicitly defined as

$$\min_{L \geq 0, R \geq 0} \frac{1}{2}\|X - LR\|_F^2 + \alpha\|L\|_F^2 + \beta \sum_{i=1}^{N} \|R(:, i)\|_1^2, \tag{8}$$

where α and β are problem-specific regularization parameters.

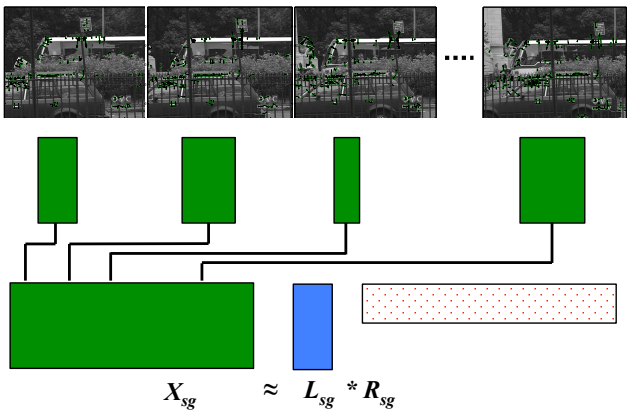


Fig. 9. Example of extracting SIFT features from a video scene and computing the compact descriptor L along with the binary selection matrix R.

B. Compact Video Scene Descriptors

Visually salient objects in a video scene maintain a nearly stationary descriptor representation throughout the GOP, resulting in significant redundancy in the columns of X. Thus, the problem of computing a compact descriptor of a video scene can be formulated as that of finding an efficient, i.e., low-dimensional, representation of the matrix X. Ideally, the set of feature vectors that represent the salient objects in a GOP can be encoded using a matrix $L \in \mathbb{R}^{m \times r}$, where $r \ll N$ represents the number of descriptors that distinctly represent the salient object. Fig. 9 illustrates the process of extracting features from a video GOP and computing the low dimensional representation L and selection matrix R.

In the case of SIFT descriptors, the columns in X are non-negative unit norm vectors. Therefore, we compute compact descriptors $\widehat{L}$ using the following alternative orthogonal non-negative matrix factorization (ONMF)

$$\begin{aligned} (\widehat{L}, \widehat{R}) = & \min_{\substack{L \in \mathbb{R}_+^{m \times r}, \\ R \in \mathbb{R}_+^{r \times N}}} \quad \tfrac{1}{2}\|X - LR\|_F^2 \\ \text{subject to} \quad & \begin{cases} \|L_i\|_2 = 1, \forall i \in \{1, \dots r\} \\ \|R_j\|_0 = 1, \forall j \in \{1, \dots N\} \end{cases}, \end{aligned} \tag{9}$$

where L_i and R_j are the columns of the matrices L and R indexed by i and j, respectively, and $\mathbb{R}_+$ is the positive orthant.

In contrast to (7), the formulation in (9) explicitly requires that only one column of L is used to represent each descriptor of X through a single non-zero coefficient in the corresponding column of R. This is equivalent to constraining RR^T to be a diagonal matrix but not necessarily the identity. Since

TABLE I
COMPRESSION RATIO OF A RANK $r = 30$ COMPACT DESCRIPTOR.

Sequence	Coastguard	Bus	Soccer	Football	Hall Monitor	Stefan
Mean descriptors per GOP	2083	6761	1055	6186	3889	11959
Compression ratio	98.66%	99.66%	97.26%	99.52%	99.33%	99.75%

L in (9) has unit norm columns, the formulation in (7) is equivalent to (9), subject to a scaling of the columns of L and, correspondingly, of R. This reformulation enables an efficient solution as described in [41].

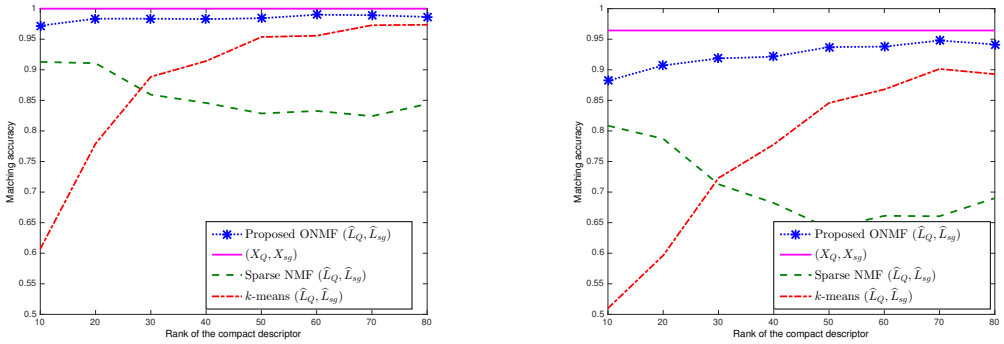
From the discussion above, it follows that the NMF formulation in (9) functions similar to a k-means classifier. For a large enough r, the columns of $\widehat{L}$ will contain the cluster centers of dominant features in the matrix X, while $\widehat{R}$ selects the cluster centers in $\widehat{L}$ that best match the data.

C. Experimental Results

We consider the problem of classifying scenes from six different video sequences. We choose the reference video sequences[1]: Coastguard, Bus, Soccer, Football, Hall Monitor, and Stefan, each composed of CIF resolution (352×288 pixels) video frames. The sequences are then divided into GOPs of size 30 frames each, and SIFT descriptors are extracted from every frame in a GOP. The sequence Stefan contains 90 frames while all other sequences contain 150 frames each. Therefore, we have a total of 28 distinct GOPs. Let s denote the video sequence index and g denote the GOP number. We stack the descriptors from GOP g of video sequence s into a matrix X_{sg} and solve the non-negative matrix factorization problem (9) to extract compact descriptors $\widehat{L}_{sg}$ with rank $r \in \{10, 20, 30, \dots 80\}$. As a representative result, Table I shows the average compression ratio per video sequence achieved by choosing a rank $r = 30$ compact descriptor.

In the scene classification experiment, our goal is to identify the video sequence to which a GOP belongs. Therefore, we choose one query GOP from the available 28 and match it to the remaining 27 database GOPs so as to classify the query GOP to a video sequence. Matching is performed by finding the GOP $\hat{g}$, whose compact descriptor $\widehat{L}_{sg}$ correlates the most with that of the query GOP $\widehat{L}_Q$. The video sequence associated with the GOP $\hat{g}$ is then chosen as the matching sequence. We also compare the matching performance of the ONMF algorithm—i.e., the algorithm that solves (9)—with that of

[1]Available from: http://trace.eas.asu.edu/yuv/



(a) (b)

Fig. 10. (a) Video scene classification accuracy using orthogonal NMF, sparse NMF, and k-means clustering for varying rank and number of clusters. (b) Classification accuracy after removing from the video database the GOPs that are temporally adjacent to the query GOP.

compact descriptors computed via k-means clustering of the SIFT features and from solving a sparse NMF problem developed in [40]. The sparse NMF formulation differs from our ONMF formulation in that the matrix R is sparse and non-binary. In all cases, the number of clusters is set equal to the rank of the matrix factors.

Fig. 10(a) illustrates the accuracy of matching a query GOP to the correct sequence using each of the three algorithms. The figure shows that compact descriptors computed using the ONMF algorithm exhibit a higher matching accuracy and are more discriminative compared to k-means or sparse NMF. Moreover, the ONMF classifier is more robust to the chosen number of clusters compared to k-means. Note that sparse NMF results in a relatively poor classifier and is very sensitive to the chosen factor rank. We also test the robustness of the compact descriptors to the scene variability by removing from the video database the GOPs that are temporally adjacent to the query GOPs. Fig. 10(b) shows the classification accuracy where the ONMF classifier maintains a superior classification performance relative to k-means and sparse NMF. Both figures demonstrate that by appropriately selecting the rank of the factorization a system designer can tune the matching accuracy vs. compression performance trade-off, according to the application constraints.

Since these methods operate on a GOP, and the size of the GOP impacts latency and buffering requirements, it is also worthwhile to study the performance with varying GOP sizes. Fig. III-C plots the video scene classification accuracy using the orthogonal NMF scheme for varying rank and GOP sizes. This plot shows that the classification performance drops less than 3% when the GOP size is changed from 30 to 5 frames, which suggests that accuracy could still be maintained under low latency constraints.

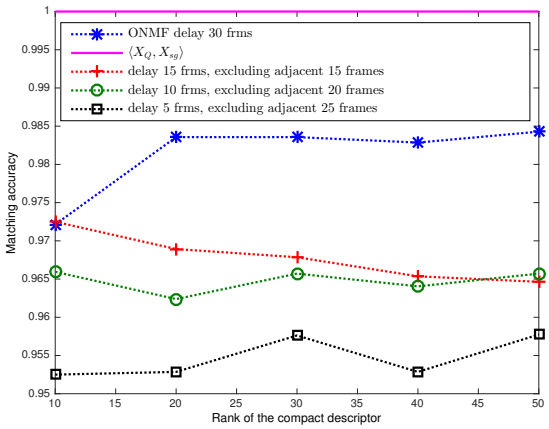


Fig. 11. Video scene classification accuracy using orthogonal NMF for varying rank and GOP sizes.

TABLE II
COMPRESSION RATIO VS. GOP SIZE AND RANK

Rank	10	20	30	40	50
30 frame GOP	99.01%	98.03%	97.04%	96.06%	95.08%
15 frame GOP	98.03%	96.06%	94.09%	92.13%	90.16%
10 frame GOP	97.04%	94.09%	91.14%	88.19%	85.24%
5 frame GOP	94.09%	88.19%	82.29%	76.39%	70.49%

Furthermore, Table II reports the change in compression ratio as a function of the rank and GOP size. These results show a reduction in the compression ratio with smaller GOP size, as one would expect. Overall, the compression benefit is still quite notable even under such constraints.

We note that combining the NMF approach with embeddings can further improve the compression efficiency of a compact descriptor. In particular, when the GOP size is large, the matrix factorization will dominate the gains. However, when the GOP size is small, the embeddings will have a greater influence on the compression of features for a particular frame. This combination might also help to reduce the complexity of the image or video matching process at the database server. The NMF approach essentially clusters the features (or embeddings) of a given image or video signal. So, while the server?s aim is still to find the matching image, its task is altered; it now has to match the query cluster against a small set of database clusters. The effectiveness of such an approach needs to be examined by further analysis.

IV. RELATED WORK

This paper has discussed two classes of dimensionality reduction approaches in the context of visual retrieval and classification. The first is based on random projections and would typically operate on

descriptors for a specific image, while the second operates over a sequence of image descriptors and uses matrix factorization or k-means clustering to identify the most salient descriptors to represent the objects in a video scene. The specific techniques described in this paper are suitable for a wide range of image/video retrieval and classification tasks. However, this is a very rapidly growing area and a number of very interesting and successful approaches have emerged in recent years. While we do not aim to provide a comprehensive review of all related methods, a select set of related techniques are discussed further in this section to provide readers with a broader sense of the available techniques that address dimensionality reduction needs in the context of visual inference problems.

In the following, we first discuss work related to random projections. Similar to the quantized embeddings presented in this paper, such techniques are independent of the data and underlying feature space, and can provide worst-case theoretical guarantees on their ability to preserve distances and achieve a specified performance. Then, we briefly touch on a few data-dependent approaches which rely on machine learning to optimize the rankings and similarities of visual data. Lastly, we review several source coding approaches including recent standardization efforts that have focused on compact descriptors for visual search.

A. Random Projections

The JL embedding and an extension to quantized embeddings has been discussed in section II. In addition to the work that has been cited, other research groups have also investigated the design and impact of quantization, such as [19], [42]–[46]. Furthermore, as noted in the paper, JL embeddings preserve ℓ_2 distances and one method to extend this to ℓ_1 distances through an isometric mapping to Hamming space has been discussed. However, there is a large body of work in preserving other similarity measurements, such as ℓ_p distances for various p's [47]–[49], edit distance [50]–[53] and the angle, i.e., correlation, between signals [54]–[56].

A common thread in the aforementioned body of work is that distances or other similarity measures are preserved indiscriminately. This is in sharp contrast to the work described in Section II-C of this paper, which allows the design of embeddings that represent some distances better than others, with control on that design. For example, in the image retrieval application, it might be beneficial to only encode a short range of distances, as necessary for nearest-neighbor computation and classification.

Recent work in this area has provided classification guarantees for JL embeddings on very particular signal models [16] and with some narrowly-defined locality properties [17]. In particular, it has been shown that separated convex ellipsoids remain separated when randomly projected to a space with

sufficient dimensions. The work described in this paper significantly enhances the available design space compared to JL embeddings. It should, thus, be possible to establish similar results, but this remains an open problem.

Another very popular and related technique is locality-sensitive hashing (LSH), which significantly reduces the computational complexity of near-neighbor computation [14], [57], [58]. The LSH literature shares many of the tools with the embeddings literature, such as randomized projections, dithering and quantization, but the goal is different: given a query point, LSH will return its near neighbors very efficiently, with $O(1)$ computation. This efficiency comes at a cost: no attempt is made to represent the distances of neighbors. When used to compare signals it only provides a binary decision, whether the distance of the signals is smaller than a threshold or not. This makes it unsuitable for applications that require more accurate distance information. That said, some of the embedding techniques could be used in the context of an LSH-based scheme. Moreover, it should also be possible to design mechanisms that reduce complexity which explicitly exploit our methods, for example extending the hierarchical approach in [59].

Two other techniques [28], [29] have been proposed for efficient remote image matching based on a version of LSH. These techniques compute random projections of scale invariant features followed by one-bit quantization based on the sign of the random projections. By construction, as the quantizer makes a 1-bit decision, these works do not consider the tradeoff between dimensionality reduction and quantization.

Finally, the work presented in [60], [61] uses randomized embeddings to efficiently approximate specific kernel computations. The application of quantized embeddings to SVM classifiers [26] is a generalization of these approaches, by allowing control over the distance map in the kernel and the ambiguity of the distance preservation.

B. Learning-Based Embeddings

There is also a large body of work focused on learning embeddings from available data. Boosting Similarity Sensitive Coding (BoostSSC) and Restricted Boltzmann Machines (RBM) have been proposed for learning compact GIST codes for content-based image retrieval [62]. Alternatively, semantic hashing can be transformed into a spectral hashing problem in which it is only necessary to calculate eigenfunctions of the GIST features, providing better retrieval performance than BoostSSC and RBM [63]. Besides these relatively recently developed machine learning algorithms, some classical training-based techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) have also been

used to generate compact image descriptors. In particular, PCA has been used to produce small image descriptors by applying techniques such as product quantization [64] and distributed source coding [65]. Alternatively, small image descriptors were obtained by applying LDA to SIFT-like descriptors followed by binary quantization [66]. Related techniques in this space include [67], [68], which attempt to learn a distance-preserving embedding from available data by solving a very expensive optimization program.

Such approaches exploit a computationally expensive training stage to improve embedding performance for a particular data set with respect to its distance-preserving properties. While significant performance gains could be realized with such approaches, the embedding guarantees are only applicable to data similar to the training data. In other words, the embedding might not generalize or perform well on different sets. For example, retraining would be required when there are significant changes or updates to the data set, or the query is performed on a data set with different content and statistics. This architecture may be undesirable for some applications, such as AR applications, in which the database can keep growing as new landmarks, new products, etc., are added.

In contrast, the approaches based on random projections are independent of the data. Such designs are considered universal in the sense that they work on any data set with overwhelming probability, as long as the embedding parameters are drawn independently of the dataset. Of course, using data for training is a promising avenue and the link between the two approaches is a very interesting area for future exploration.

C. Source Coding of Descriptors

As a source coding-based alternative to random projection methods and learning-based dimensionality reduction, a low-bitrate descriptor has been constructed using a Compressed Histogram of Gradients (ChoG) specifically for augmented reality applications [69]. In this method, gradient distributions are explicitly compressed, resulting in low-rate scale invariant descriptors.

Along these lines, a new standard referred to as Compact Descriptor for Visual Search (CDVS) has been recently finalized to provide an interoperable solution for state-of-the-art image-based retrieval [70]. The main steps employed in the extraction and encoding pipeline of this standard are as follows:

- Keypoint detection: To handle the scale invariance, keypoints are identified based on the creation of a scale-space made by a set of Laplacian-of-Gaussian (LoG) filtered images and the subsequent identification of extrema in this space by means of polynomial approximations.
- Feature selection: Based on the characteristics and relevance of the keypoints, a subset of keypoints is extracted, so as to maximize a measure of expected quality for subsequent matching.

- Local descriptor compression: To reduce the bit rate, transform and scalar quantization-based compression techniques are applied to selected local descriptors.
- Coordinate coding: An independent compression of the coordinates of the selected key points, which are critical for geometric verification, is additionally applied.
- Global descriptor aggregation: The local descriptors are finally aggregated to form a single global descriptor. This stage includes dimensionality reduction through PCA and formation of a Scalable Fisher Vector which is then binarized.

Further details on the above steps and performance comparisons relative to other image-based descriptors can be found in [70].

In order to determine an efficient representation of image descriptors over a sequence of images, temporal correlation among descriptors should also be considered. The approach discussed in section III essentially summarizes descriptors to a small set that can represent the visually salient objects in the video scene. We discussed the use of matrix factorization or k-means clustering techniques for this purpose.

Alternatively, the sequence of image descriptors could be compressed using traditional source coding techniques that exploit the motion of those descriptors through the sequence. Examples of such techniques appear in [71]–[74]. These methods exploit powerful paradigms from video compression, such as motion compensated prediction and rate-distortion optimization, to reduce the bit-rate of the transmitted descriptors.

In terms of future standardization, MPEG is also planning to develop a standard for Compact Descriptors for Visual Analysis (CDVA) [75]. It is probable that the new standard will extend the CDVS standard from images to video signals, and that it will employ techniques similar to those discussed in this paper and above.

V. Concluding Remarks

Advanced media applications such as augmented reality and situation-aware systems will be enabled through techniques that efficiently perform inference algorithms on visual data. Much work has been done in recent years to identify features that represent the salient aspects of the visual information and facilitate a wide range of inference tasks, including similarity search, semantic indexing and classification. This paper has reviewed several relevant requirements for such systems, including the need to communicate visual features with low rate and latency, and facilitate inference with low complexity.

In this paper we argue that dimensionality reduction is critical technology to satisfy the needs of these systems. In particular, we review two types of schemes for dimensionality reduction: quantized embed-

dings, which offer the ability to preserve distances and satisfy rate constraints in a lower dimensional space, and a matrix factorization approach, which summarizes the most relevant features to describe the sequence of descriptors associated with a video scene. Both approaches enable noteworthy rate savings in visual inference applications and provide significant flexibility in navigating the rate vs. performance trade-off, similar to the rate-distortion trade-off in conventional compression.

The specific methods presented in this paper are part of a much larger body of work that addresses dimensionality reduction techniques for visual applications. For instance, many recent works have shown the benefits of learning low-dimensional embeddings to optimize similarity search. Also, the standardization of compact descriptors for visual search and analysis is underway. One standard addressing the needs for image data is already complete while plans for future standards that extend these approaches to video are actively being discussed and considered.

We hope that the material presented in this paper illustrates some of the emerging applications that require advanced processing of visual data, and highlights relevant technology in the current literature. Beyond this, we believe that there are opportunities for new methods that efficiently represent and encode visual information for more general functions, such as classifiers and estimators based on machine learning algorithms. We also expect to see applications in distributed environments, where visual information may be partially observed and processed in a decentralized manner on nodes in a network, with the goal of performing joint inference or control.

REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, pp. 91–110, 2004.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, 2008.

[3] 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.

[4] 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.

[5] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 510–517.

[6] J. Sivic and A. Zisserman, "Efficient visual search of videos cast as text retrieval," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 591–606, Apr. 2009.

[7] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615 –1630, Oct. 2005.

[8] M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of cognitive neuroscience*, vol. 3, no. 1, pp. 71–86, 1991.

[9] M. Jones and P. Viola, "Face recognition using boosted local features," *MERL Technical Report, TR2003-25*, May 2003.

[10] A. K. Jain, P. J. Flynn, and A. A. Ross, *Handbook of biometrics*. Springer, 2008.

[11] H. Lee, R. Grosse, R. Ranganathan, and A. Ng, "Proceedings of the 26th annual international conference on machine learning," in *Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations*, 2009, pp. 609–616.

[12] W. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189 –206, 1984.

[13] S. Dasgupta and A. Gupta, "An elementary proof of a theorem of Johnson and Lindenstrauss," *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

[14] P. Indyk and R. Motwani, "Approximate nearest neighbors: towards removing the curse of dimensionality," in *ACM Symposium on Theory of computing*, 1998, pp. 604–613.

[15] P. T. Boufounos, S. Rane, and H. Mansour, "Representation and coding of signal geometry," *arXiv preprint arXiv:1512.07636*, 2015.

[16] A. S. Bandeira, D. G. Mixon, and B. Recht, "Compressive classification and the rare eclipse problem," *arXiv preprint arXiv:1404.3203*, 2014.

[17] S. Oymak and B. Recht, "Near-optimal bounds for binary embeddings of arbitrary sets," *arXiv preprint arXiv:1512.04433*, 2015.

[18] M. Li, S. Rane, and P. Boufounos, "Quantized embeddings of scale-invariant image features for mobile augmented reality," in *IEEE International Workshop on Multimedia Signal Processing*, Banff, Canada, September 17-19 2012.

[19] L. Jacques, "Small width, low distortions: quasi-isometric embeddings with quantized sub-gaussian random projections," *arXiv preprint arXiv:1504.06170*, 2015.

[20] P. Indyk, "Stable distributions, pseudorandom generators, embeddings, and data stream computation," *Journal of ACM*, vol. 53, no. 3, pp. 307–323, 2006.

[21] W. Lu, A. Varna, and M. Wu, "Secure Image Retrieval through Feature Detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Taipei, Taiwan, 2009.

[22] N. Linial, E. London, and Y. Rabinovich, "The geometry of graphs and some of its algorithmic applications," *Combinatorica*, vol. 15, no. 2, pp. 215–245, 1995. [Online]. Available: http://dx.doi.org/10.1007/BF01200757

[23] S. Rane, P. Boufounos, and A. Vetro, "Quantized embeddings: An efficient and universal nearest neighbor method for cloud-based image retrieval," in *SPIE Optics and Photonics*, San Diego, CA, August 25-29 2013.

[24] P. T. Boufounos and S. Rane, "Secure binary embeddings for privacy preserving nearest neighbors," in *Proc. Workshop on Information Forensics and Security (WIFS)*, November 29 - December 2 2011. [Online]. Available: http://dx.doi.org/10.1109/WIFS.2011.6123149

[25] S. Rane and P. T. Boufounos, "Privacy-preserving nearest neighbor methods: Comparing signals without revealing them," *IEEE Signal Processing Magazine*, March 2013. [Online]. Available: http://dx.doi.org/10.1109/MSP.2012.2230221

[26] P. T. Boufounos and H. Mansour, "Universal embeddings for kernel machine classification," in *Proceedings of the International Conference on Sampling Theory and Applications (SampTA)*, 2015, pp. 307–311.

[27] H. Shao, T. Svoboda, and L. V. Gool, "ZuBuD : Zurich Buildings database for image based recognition," Computer Vision Lab, Swiss Federal Institute of Technology, Switzerland, Tech. Rep. 260, Apr. 2003. [Online]. Available: http://www.vision.ee.ethz.ch/showroom/zubud/

[28] C. Yeo, P. Ahammad, and K. Ramchandran, "Rate-efficient visual correspondences using random projections," in *IEEE International Conference on Image Processing*, Oct. 2008, pp. 217 –220.

[29] K. Min, L. Yang, J. Wright, L. Wu, X.-S. Hua, and Y. Ma, "Compact projection: Simple and efficient near neighbor search with practical memory requirements," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, Jun. 2010, pp. 3477 –3484.

[30] D. Navneet and B. Triggs, "Histograms of oriented gradients for human detection," in *International Conference on Computer Vision & Pattern Recognition*, vol. 2, June 2005, pp. 886–893.

[31] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[32] "INRIA Person Dataset," http://pascal.inrialpes.fr/data/human/.

[33] L. Fei-Fei, R. Fergus, and P. Perona, "Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories," in *Proc. IEEE Conf. on Comp. Vision and Pattern Recognition (CVPR), Workshop on Generative-Model Based Vision.*, June 2004, pp. 178–178.

[34] "Caltech 101 dataset," http://www.vision.caltech.edu/Image_Datasets/Caltech101/.

[35] P. T. Boufounos and S. Rane, "Efficient coding of signal distances using universal quantized embeddings," in *Proc. Data Compression Conference (DCC)*, Snowbird, UT, March 20-22 2013.

[36] P. Paatero and U. Tapper, "Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[37] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, p. 788, october 1999.

[38] N. Gillis, "The why and how of nonnegative matrix factorization," *http://arxiv.org/abs/1401.5226*, 2014.

[39] C. Ding, T. Li, W. Peng, and H. Park, "Orthogonal nonnegative matrix t-factorizations for clustering," in *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '06, 2006, pp. 126–135.

[40] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, jun 2007.

[41] H. Mansour, S. Rane, P. Boufounos, and A. Vetro, "Video querying via compact descriptors of visually salient objects," in *Proceedings of IEEE International Conference on Image Processing*, 2014.

[42] L. Jacques, "A quantized johnson-lindenstrauss lemma: The finding of buffon's needle," *IEEE Trans. Info. Theory*, vol. 61, no. 9, pp. 5012–5027, Sept 2015.

[43] L. Jacques, J. N. Laska, P. T. Boufounos, and R. G. Baraniuk, "Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors," *IEEE Trans. Info. Theory*, vol. 59, no. 4, April 2013. [Online]. Available: http://dx.doi.org/10.1109/TIT.2012.2234823

[44] Y. Plan and R. Vershynin, "One-bit compressed sensing by linear programming," *Communications on Pure and Applied Mathematics*, vol. 66, no. 8, pp. 1275–1297, 2013.

[45] ——, "Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach," *Information Theory, IEEE Transactions on*, vol. 59, no. 1, pp. 482–494, 2013.

[46] ——, "Dimension reduction by random hyperplane tessellations," *Discrete & Computational Geometry*, vol. 51, no. 2, pp. 438–461, 2014.

[47] P. Indyk, "Stable distributions, pseudorandom generators, embeddings, and data stream computation," *Journal of the ACM (JACM)*, vol. 53, no. 3, pp. 307–323, 2006.

[48] L. Jacques, D. K. Hammond, and J. M. Fadili, "Dequantizing compressed sensing: When oversampling and non-gaussian constraints combine," *Information Theory, IEEE Transactions on*, vol. 57, no. 1, pp. 559–571, 2011.

[49] L. Jacques, D. K. Hammond, and M.-J. Fadili, "Stabilizing nonuniformly quantized compressed sensing with scalar companders," *IEEE Trans. Info. Theory*, vol. 59, no. 12, pp. 7969–7984, 2013.

[50] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, Feb. 1966.

[51] A. Andoni, M. Deza, A. Gupta, P. Indyk, and S. Raskhodnikova, "Lower bounds for embedding edit distance into normed spaces," in *Proceedings of the fourteenth annual ACM-SIAM symposium on Discrete algorithms*, 2003, pp. 523–526.

[52] Z. Bar-Yossef, T. Jayram, R. Krauthgamer, and R. Kumar, "Approximating edit distance efficiently," in *Foundations of Computer Science, 2004. Proceedings. 45th Annual IEEE Symposium on*. IEEE, 2004, pp. 550–559.

[53] R. Ostrovsky and Y. Rabani, "Low distortion embeddings for edit distance," *Journal of the ACM (JACM)*, vol. 54, no. 5, p. 23, 2007.

[54] P. T. Boufounos, "Sparse signal reconstruction from phase-only measurements," in *Proc. Int. Conf. Sampling Theory and Applications (SampTA)*, Bremen, Germany, July 1-5 2013.

[55] ——, "On embedding the angles between signals," in *Proc. Signal Processing with Adaptive Sparse Structured Representations (SPARS)*, Lausanne, Switzerland, July 8-11 2013.

[56] ——, "Angle-preserving quantized phase embeddings," in *Proc. SPIE Wavelets and Sparsity XV*, San Diego, CA, August 25-29 2013.

[57] M. Datar, N. Immorlica, P. Indyk, and V. S. Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *Proceedings of the twentieth annual symposium on Computational geometry*, ser. SCG '04. New York, NY, USA: ACM, 2004, pp. 253–262.

[58] A. Andoni and P. Indyk, "Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions," *Commun. ACM*, vol. 51, no. 1, pp. 117–122, Jan. 2008.

[59] P. T. Boufounos, "Hierarchical distributed scalar quantization," in *Proc. Int. Conf. Sampling Theory and Applications (SampTA)*, Singapore, May 2-6 2011. [Online]. Available: http://sampta2011.ntu.edu.sg/SampTA2011Proceedings/papers/Tu5S06.4-P0226.pdf

[60] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in *Advances in neural information processing systems*, 2007, pp. 1177–1184.

[61] M. Raginsky and S. Lazebnik, "Locality-sensitive binary codes from shift-invariant kernels," *The Neural Information Processing Systems*, vol. 22, 2009.

[62] A. Torralba, R. Fergus, and Y. Weiss, "Small codes and large image databases for recognition," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, Jun. 2008, pp. 1 –8.

[63] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Advances in Neural Information Processing Systems 21*, D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Eds., 2009, pp. 1753–1760.

[64] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating local images descriptors into compact codes," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, p. 1, 2011.

[65] C. Yeo, P. Ahammad, and K. Ramchandran, "Coding of image feature descriptors for distributed rate-efficient visual correspondences," *International Journal of Computer Vision*, vol. 94, pp. 267–281, 2011, 10.1007/s11263-011-0427-1.

[66] C. Strecha, A. Bronstein, M. Bronstein, and P. Fua, "LDAHash: Improved matching with smaller descriptors," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 66 –78, Jan. 2012.

[67] C. Hegde, A. Sankaranarayanan, W. Yin, and R. Baraniuk, "NuMax: A convex approach for learning near-isometric linear embeddings," *IEEE Trans. Signal Processing*, vol. 63, no. 22, pp. 6109–6121, Nov 2015.

[68] A. Sadeghian, B. Bah, and V. Cevher, "Energy-aware adaptive bi-Lipschitz embeddings," in *Proc. Int. Conf. Sampling Theory and Applications (SampTA)*, Bremen, Germany, July 1-5 2013.

[69] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, Y. Reznik, R. Grzeszczuk, and B. Girod, "Compressed histogram of gradients: A low-bitrate descriptor," *International Journal of Computer Vision*, vol. 96, pp. 384–399, 2012.

[70] L.-Y. Duan, J. Lin, J. Chen, T. Huang, and W. Gao, "Compact descriptors for visual search," *IEEE Multimedia*, vol. 21, no. 3, pp. 30 – 40, 2014.

[71] L. Baroffio, M. Cesana, A. Redondi, S. Tubaro, and M. Tagliasacchi, "Coding video sequences of visual features," in *IEEE International Conference on Image Processing (ICIP 2013)*, Melbourne, Australia, Sep. 2013.

[72] M. Makar, S. Tsai, V. Chandrasekar, D. Chen, and B. Girod, "Interframe coding of canonical patches for low bit-rate mobile augmented reality," *International Journal of Semantic Computing*, vol. 7, no. 01, pp. 5–24, 2013.

[73] M. Makar, S. Tsai, V. Chandrasekhar, D. Chen, and B. Girod, "Interframe coding of canonical patches for mobile augmented reality," in *Multimedia (ISM), 2012 IEEE International Symposium on*. IEEE, 2012, pp. 50–57.

[74] L. Baroffio, J. Ascenso, M. Cesana, A. Redondi, and M. Tagliasacchi, "Coding binary local features extracted from video sequences," in *IEEE International Conference on Image Processing*, 2014.

[75] MPEG Requirements, "Call for Proposals for Compact Descriptors for Video Analysis (CDVA) – Search and Retrieval," *ISO/IEC JTC1/SC29/WG11, Doc N15339*, July 2015.

Biographies

Petros T. Boufounous is Senior Principal Researcher and the Sensing Team Leader at Mitsubishi Electric Research Laboratories (MERL), and a visiting scholar at the Rice University ECE department. Dr. Boufounos completed his undergraduate and graduate studies at MIT. He received the S.B. degree in Economics in 2000, the S.B. and M.Eng. degrees in Electrical Engineering and Computer Science in 2002, and the Sc.D. degree in EECS in 2006. Between September 2006 and December 2008, he was a postdoctoral associate with the DSP Group at Rice University. His research focus includes signal acquisition and processing theory, quantization, and data representations. He is also interested in interactions with fields that use sensing extensively, such as machine learning and robotics. Dr. Boufounos is a Senior Area Editor at IEEE Signal Processing Letters. He is a senior member of the IEEE and a member of Sigma Xi, Eta Kappa Nu, and Phi Beta Kappa.

Hassan Mansour is a Principal Research Scientist in the Multimedia Group at Mitsubishi Electric Research Laboratories, Cambridge, MA. He received his M.A.Sc. (2005) and Ph.D. (2009) from the Department of Electrical and Computer, University of British Columbia (UBC), Vancouver, Canada where he conducted research on scalable video coding and transmission. He then pursued a postdoctoral fellowship in the Departments of Mathematics, Computer Science, and Earth and Ocean Sciences at UBC working on theoretical and algorithmic aspects of compressed sensing and its application to seismic imaging. His current research includes developing algorithms for video analytics as well as for sensing parsimonious signals.

Shantanu Rane is a Senior Member of the Research Staff at Xerox PARC. He has a M.S. degree from The University of Minnesota (2001) and a Ph.D. degree from Stanford University (2007), both in Electrical Engineering. From 2007-2014, he worked at Mitsubishi Electric Research Laboratories (MERL) in Cambridge, MA. His research interests are in the areas of applied cryptography, information theory and statistical signal processing. He has participated in standardization activity for the Joint Video Team (JVT) within the ITU-T/MPEG H.264/AVC video compression standard; INCITS-M1, the US National Body for standardization of biometrics; and the ISO/IEC JTC1 SC37 Subcommittee on Biometrics. He currently serves as an Associate Editor for the IEEE Signal Processing Magazine.

Anthony Vetro is a Deputy Director at Mitsubishi Electric Research Labs, in Cambridge, MA. He also manages a group that is responsible for research in the areas of digital video coding and processing, information security, sensing technologies, and speech/audio processing. He has contributed to the transfer and development of several technologies to Mitsubishi products, including digital television receivers and displays, surveillance and camera monitoring systems, automotive equipment, as well as satellite imaging systems. He has published more than 200 papers and has been an active member of the MPEG and ITU-T video coding standardization committees for a number of years. He is also active in various IEEE conferences, technical committees and editorial boards. Dr. Vetro received the B.S., M.S. and Ph.D. degrees in Electrical Engineering from Polytechnic University, in Brooklyn, NY, and is a Fellow of the IEEE.