VIDEO QUERYING VIA COMPACT DESCRIPTORS OF VISUALLY SALIENT OBJECTS

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

Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, USA
{mansour, rane, petrosb, avetro}@merl.com

ABSTRACT

We consider the problem of extracting descriptors that represent visually salient portions of a video sequence. Most state-of-the-art schemes generate video descriptors by extracting features, e.g., SIFT or SURF or other keypoint-based features, from individual video frames. This approach is wasteful in scenarios that impose constraints on storage, communication overhead and on the allowable computational complexity for video querying. More importantly, the descriptors obtained by this approach generally do not provide semantic clues about the video content. In this paper, we investigate new feature-agnostic approaches for efficient retrieval of similar video content. We evaluate the efficiency and accuracy of retrieval when k-means clustering is applied to image features extracted from video frames. We also propose a new approach in which the extraction of compact video descriptors is cast as a Non-negative Matrix Factorization (NMF) problem. Initial experiments on video-based matching suggest that compact descriptors obtained via low-rank matrix factorization provide a better combination of discriminability and robustness to rank variations than those obtained via k-means clustering.

Index Terms— k-means, NMF, descriptors, video retrieval

1. INTRODUCTION

The advent of inexpensive cameras and inexpensive storage has made practical the collection and storage of large databases of images or video sequences. The commercial viability of such databases depends in large part on the availability of search and retrieval tools. Thus, much research activity has been devoted to retrieval mechanisms for images. In general, such mechanisms rely on identifying points of interest in an image, often referred to as keypoints, and then extracting features from these points that are robust to variations in translation, rotation, scaling and illumination. Examples of such features include SIFT [1], SURF [2], BRISK [3], FREAK [4], HoG [5], CHoG [6] and others. To reduce the bandwidth and complexity while preserving matching accuracy and speed, the features are often aggregated and summarized to more compact descriptors. Approaches for compacting the feature spaces include Principal Component Analysis (PCA) [7], Linear Discriminant Analysis (LDA) [8], Boosting [9], Spectral hashing [10], and the popular Bag-of-Features approach [11]. The latter converts features to compact descriptors (codewords) using the cluster centers produced by k-means clustering. The compact descriptors extracted from a query image, are then compared to those extracted from images in the database in order to determine similar images. There has, however, been much less work in developing efficient feature matching mechanisms for video queries.

Extending existing image descriptors to derive video descriptors is not straightforward. One naive approach would be to extract image descriptors from each frame in the video sequence, treating frames as separate images. This approach fails to exploit the fact that features extracted from successive video frames will be very similar and describe similar keypoints, resulting in a very redundant representation. Furthermore, it does not remove features that are not persistent from frame to frame and probably do not describe the video sequence very well. Thus, simply collecting individual image descriptors would be bandwidth-inefficient and would significantly increase matching complexity. A vastly more efficient approach is to compress the descriptors derived from each video frame, exploiting the motion of those descriptors through the sequence [12–14]. 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. They do not, however, address the problem of discovering a small set of descriptors that can represent the visually salient object.

In this paper, we investigate new approaches to video-based retrieval using compact descriptors. While we seek to leverage the extensive prior work on image descriptors for particular applications, our goal is to provide a general (feature-agnostic) mechanism to extend image descriptors to compact video descriptors. Hence, our experimental results will use SIFT descriptors, but our approach is universally applicable for retrieval based on nearest-neighbor search. Our first step is to evaluate the transmission efficiency and retrieval accuracy achieved by performing k-means clustering of image descriptors extracted from a video sequence. It is well-known that k-means gives a locally optimal clustering that may be sensitive to the initial conditions. As an alternative to performing k-means in the traditional way, we propose a principled approach to the extraction of compact video descriptors based on Non-negative Matrix Factorization (NMF). There is a close fundamental relationship between the rank of the descriptor matrix obtained via NMF and the number of clusters in the k-means algorithm [15]. Nevertheless, our experiments suggest that compact descriptors obtained via the NMF approach are more discriminative than those obtained via traditional k-means, while also ensuring that the retrieval accuracy is much less sensitive to the chosen number of clusters.

The following section provides some background on NMF and establishes notation. Section 3 describes the proposed approach which formulates the derivation of compact video descriptors as a NMF problem. The proposed approach is experimentally validated in Section 4. We discuss our results and conclude in Section 5.

2. NON-NEGATIVE MATRIX FACTORIZATION

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

$$X \approx LR. \tag{1}$$

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 [16], NMF has gained popularity in machine learning and data mining following the work of Lee and Seung [17]. 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 [18].

Of interest for this paper are NMF formulations proposed for clustering [19, 20]. Specifically, we consider the sparse NMF and orthogonal NMF formulations. The orthogonal NMF problem is defined as

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

which was shown in [19] to be equivalent to k-means clustering. Alternatively, the sparse NMF problem [20] 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

$$\min_{L\geq 0, R\geq 0} \frac{1}{2}\|X - LR\|_F^2 + \alpha\|L\|_F^2 + \beta\sum_{i=1}^{N}\|R(:,i)\|_1^2, \tag{3}$$

where α and β are problem specific regularization parameters.

Note that NMF problems are non-convex; algorithms that tackle these problems generally do not have global optimality guarantees. Therefore, different algorithms that tackle the same problem may arrive at different solutions. In what follows, we develop an algorithm that addresses the orthogonal NMF problem and demonstrate that the solutions produced by our algorithm enjoy better classification properties compared to k-means and sparse NMF.

3. COMPACT SCENE DESCRIPTORS

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. Our framework assumes 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. In this section, we develop a framework for extracting a compact descriptor that represents the salient visual information in a video scene.

3.1. Computing the Compact Descriptor Using NMF

We observe that visually salient objects in a video scene maintain a nearly stationary descriptor representation throughout the GOP. Therefore, we formulate the problem of computing a compact descriptor of a video scene as that of finding a 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. 1 illustrates the process of extracting features from a video GOP and computing the low dimensional representation L and selection matrix R.

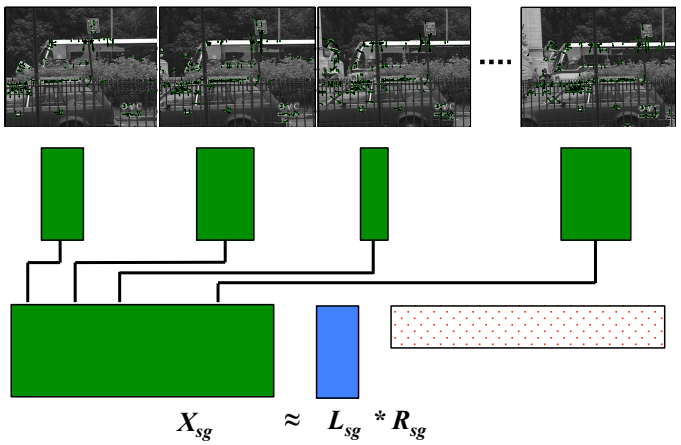


Fig. 1: Example of extracting SIFT features from a video scene and computing the compact descriptor L along with the binary selection matrix R.

In the case of SIFT descriptors, the columns in X are non-negative unit norm vectors. Therefore, we pose the problem of computing $\widehat{L}$ as the following non-negative matrix factorization (NMF) problem

$$\begin{aligned} \min_{\substack{L\in\mathbb{R}_+^{m\times r},\\ R\in\mathbb{R}_+^{r\times N}}} \quad & \tfrac{1}{2}\|X - LR\|_F^2 \\ \text{subject to} \quad & \begin{cases} \|L_i\|_2 = 1, \forall i \in \{1,\dots r\} \\ \|R_j\|_0 = 1, \forall j \in \{1,\dots N\} \end{cases}, \end{aligned} \tag{4}$$

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.

The NMF formulation in (4) functions similar to a k-means classifier and ensures that 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. In order to solve (4), we develop the projected proximal-point alternating least squares minimization algorithm shown in Algorithm 1. In every iteration k of the algorithm, the factors L_k and R_k are updated by first finding the minimizer of the proximal least squares terms

$$\begin{aligned} \widetilde{L} &= \arg\min_{L\in\mathbb{R}^{m\times r}} \tfrac{1}{2}\|X - LR_k\|_F^2 + \tfrac{\rho}{2}\|L - L_k\|_F^2, \\ \widetilde{R} &= \arg\min_{R\in\mathbb{R}^{r\times N}} \tfrac{1}{2}\|X - L_kR\|_F^2 + \tfrac{\rho}{2}\|R - R_k\|_F^2. \end{aligned} \tag{5}$$

The columns of $\widetilde{L}$ are then projected onto the non-negative ℓ_2 unit ball, while the columns of $\widetilde{R}$ are projected onto the admissible set of standard basis vectors

$$E_r := \{e_i \in \mathbb{R}^r : e(i) = 1, \text{ and } 0 \text{ otherwise}, \ i \in \{1,\dots r\}\}$$

by setting the largest non-negative entry in each column to one and the remaining entries to zero. Note that $\widetilde{L}$ and $\widetilde{R}$ admit closed-form solutions as shown in Algorithm 1. The factors L_0 and R_0 are initialized with independent identically distributed uniform random entries. The iterates $\widetilde{L}$ and $\widetilde{R}$ are computed by solving proximal-point alternating least squares functionals and then keeping only the pos-

itive entries $\widetilde{L}_+$ and $\widetilde{R}_+$ in the factors. Finally, the factors are projected onto the unit column norm ball for $\widetilde{L}$, and onto the binary selector set E_r for $\widetilde{R}$.

Algorithm 1 Projected Proximal-point Alternating Least Squares

Input X, factor rank r, ρ, `maxiter`
Output $\widehat{L}$, $\widehat{R}$
Initialize $k = 0$, $L_0 \in U_{[0,1]}$, $R_0 \in U_{[0,1]}$
while not converged and $k <$ `maxiter` **do**
  Update the compact descriptor L
  $\widetilde{L} = \left(\rho L_k + X R_k^T\right)\left(\rho I_r + R_k R_k^T\right)^{-1}$
  $L_{k+1,i} = \widetilde{L}_{+i} / \|\widetilde{L}_{+i}\|_2$
  Update the binary selector R
  $\widetilde{R} = \left(\rho I_r + L_k^T L_k\right)^{-1}\left(\rho R_k + L_{k+1}^T X\right)$
  $R_{k+1,j} = \text{Proj}_{E_r}(\widetilde{R}_j),\ \forall j \in \{1, \dots n\}$
  $k = k + 1$
end while
$\widehat{L} = L_k$, $\widehat{R} = R_k$

3.2. Classification Using Compact Descriptors

Consider the problem of classifying a video scene with respect to a database of video sequences. By extracting compact descriptors $\widehat{L}$ from video GOPs, we can now reduce the problem of matching all feature vectors in a query GOP with the features in the video database to that of matching the compact descriptors between the query GOP and the database GOPs.

Suppose that the query video as well as the database videos are divided into GOPs of size n video frames. Let $\widehat{L}_Q$ denote the query GOP's compact descriptor and $\widehat{L}_D(g)$ denote the compact descriptors of the database GOPs indexed by g. We say that a database GOP indexed by $\hat{g}$ matches the query GOP if it has the largest correlation coefficient relative to $\widehat{L}_Q$, i.e.

$$\hat{g} = \arg\max_g \|\widehat{L}_Q^T \widehat{L}_D(g)\|_\infty, \tag{6}$$

where the infinity norm $\|\cdot\|_\infty$ is applied after vectorizing the matrix product $\widehat{L}_Q^T \widehat{L}_D(g)$. Consequently, the matching GOP is the one whose compact descriptor correlates most with the query descriptor.

4. 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 composed of CIF resolution (352×288 pixels) video frames and shown in Fig. 2. 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. We stack the descriptors from GOP g of video sequence s into a matrix X_{sg} and solve the non-negative matrix factorization problem (4) using Algorithm 1 to extract compact descriptors $\widehat{L}_{sg}$ with rank $r \in \{10, 20, 30, \dots 80\}$. As a representative result, Table 1 shows the average compression ratio per video sequence achieved by choosing a rank $r = 30$ compact descriptor.

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

Table 1: Compression ratio of a rank $r = 30$ compact descriptor.

Sequence	Coast-guard	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%

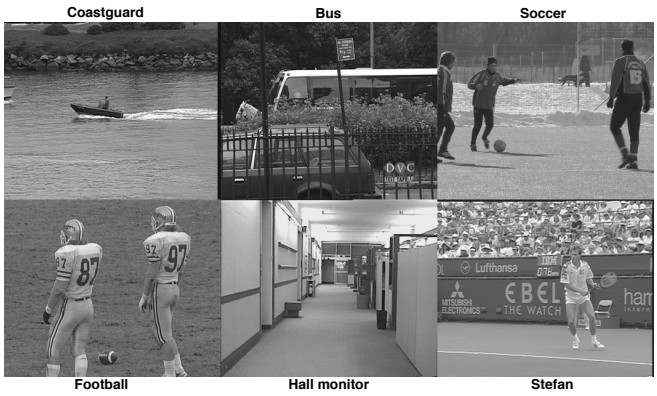


Fig. 2: First frame from each of the six reference video sequences Coastguard, Bus, Soccer, Football, Hall monitor, and Stefan.

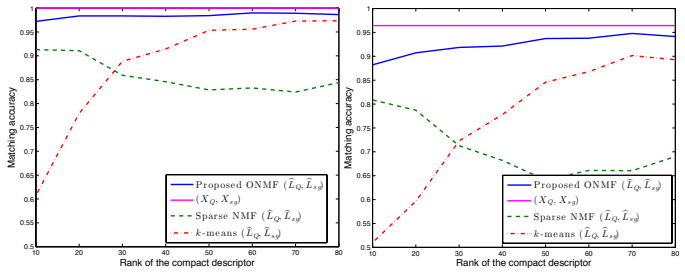
4.1. Scene classification

In the scene classification experiment, we wish to identify the video 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 according to (6) by finding the GOP 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 our ONMF algorithm with that of compact descriptors computed via k-means clustering of the SIFT features and from solving a sparse NMF problem developed in [20]. 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. 3(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. 3(b) shows the classification accuracy where the ONMF classifier maintains its superior classification performance relative to k-means and sparse NMF.

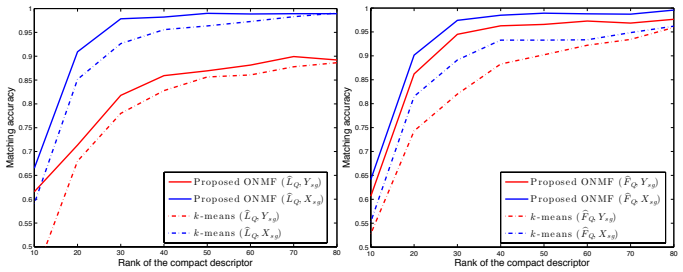
4.2. Object detection

In the object detection experiment, we wish to classify a moving object in a dynamic video scene by matching it to a video database. We assume that a user device captures a query video GOP and computes



(a) (b)

Fig. 3: (a) Video scene classification accuracy using orthogonal NMF, sparse NMF, and k-means clustering for varying factor rank and number of clusters. (b) Classification accuracy after removing from the video database the GOPs that are temporally adjacent to the query GOP.



(a) (b)

Fig. 4: (a) Object detection accuracy using orthogonal NMF and k-means clustering for varying factor rank and number of clusters, where query descriptors extracted from the complete scene are matched against database descriptors extracted either from complete scenes or just from salient objects. (b) Object detection accuracy using orthogonal NMF and k-means clustering for varying factor rank and number of clusters, where query descriptors extracted only from salient objects are matched against database descriptors extracted either from complete scenes or just from salient objects.

a compact descriptor for the GOP before transmitting it for matching in a video database. We consider two scenarios where (1) the video database contains complete visual scenes and (2) the database contains videos of separated salient objects. The visually salient objects can be extracted from dynamic video scenes via background subtraction as in [21].

Let $\widehat{L}_Q$ and $\widehat{F}_Q$ denote the compact descriptors extracted from the SIFT features of the complete query scene and the SIFT features of only the salient objects in the query scene, respectively. Also, denote by X_{sg} and Y_{sg} the SIFT features in the video database of sequence s and GOP g for the complete video scenes and for the salient objects in the video scene, respectively.

Fig. 4(a) shows the accuracy of matching the compact descriptor $\widehat{L}_Q$ to the database descriptors X_{sg} and Y_{sg}. The performance degrades significantly when $\widehat{L}_Q$ is matched with the salient features alone. However, when the salient objects are segmented before computing the compact descriptor $\widehat{F}_Q$, Fig. 4(b) shows that the matching performance is almost similar to the $(\widehat{L}_Q, X_{sg})$ combination, i.e., matching descriptors extracted from the complete video scenes. Moreover, the matching accuracy of $\widehat{F}_Q$ is only mildly affected by the exclusion of non-salient features in the database content. Finally, we note that the ONMF solution outperforms the k-means solution irrespective of whether the matching is done using features extracted from the entire video frame or using features extracted only from salient objects in the video scene.

5. DISCUSSION AND CONCLUSION

The aim of these experiments is to highlight the benefit of summarizing the feature space for reducing both the query size and the storage requirements in a video database. Our experiments demonstrate that low dimensional clustering of visual features can significantly reduce the memory requirements for representing visually salient objects in a video scene. The results in Table 1 show that a rank 30 compact descriptor achieves storage reductions that exceed 97% and average at 99%. Moreover, the compact descriptors maintain their discriminability with well over 90% matching accuracy despite the significant compression.

Algorithmically, we demonstrate that our proposed orthogonal NMF (ONMF) method for finding low dimensional clusters is more discriminative than both k-means clustering and sparse NMF. Our approach is also more robust to variations in the number of clusters than k-means. One striking observation is that while sparse NMF outperforms k-means for very low-dimensional compact representations, it quickly becomes unstable as the number of clusters, i.e., the rank of the factors, increases. We note here that since all of the above mentioned clustering problems are non-convex, the solutions to these problems depend on the initialization. Therefore, every point in our plots is an average over 50 trials of running each algorithm.

The second set of experiments on object detection highlight the effect of restricting the compact representation to features extracted only from visually salient objects in a scene. The experiments show that if the video database contains descriptors of visually salient objects alone, then computing a compact descriptor of the full query scene negatively impacts the matching accuracy. On the other hand, restricting the compact representation to the visually salient object features enjoys high matching accuracy for both cases where the database contains full scene features or visually salient object features alone.

It is still unclear to us why our proposed ONMF algorithm performs better than standard k-means clustering. It may be that the smoothing induced by the proximal-point approach helps in avoiding local minima in which k-means tends to get stuck. This remains an open question to be resolved in future work.

In conclusion, we have shown that feature clustering is quite successful at extracting discriminative representations of high-dimensional feature spaces while significantly reducing the storage and transmission requirements. Particularly in the case of video scenes, compact descriptors computed via low-rank non-negative matrix factorization can exploit the near stationarity of salient object descriptors in the scene. We developed an efficient algorithm that finds low dimensional clusters from the deluge of visual descriptors extracted from videos. We also demonstrated through experimental validation that compact descriptors extracted using our proposed approach provide better discriminability and are more robust to rank variations than those obtained by k-means clustering and sparse NMF.

6. 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. Van Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346 – 359, 2008.

[3] 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.

[4] 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.

[5] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, vol. 1, pp. 886–893.

[6] 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.

[7] 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. 34, no. 9, pp. 1704–1716, Sept 2012.

[8] C. Strecha, A.M. Bronstein, M.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.

[9] 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*, June 2008, pp. 1 –8.

[10] 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., pp. 1753–1760. 2009.

[11] 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.

[12] 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, Sept. 2013.

[13] 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.

[14] 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.

[15] J. Kim and H. Park, "Toward faster nonnegative matrix factorization: A new algorithm and comparisons," in *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08)*, 2008, pp. 353–362.

[16] 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.

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

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

[19] 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*, 2006, KDD '06, pp. 126–135.

[20] 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.

[21] H. Mansour and A. Vetro, "Video background subtraction using semi-supervised robust matrix completion," in *to appear in Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.