VIDEO CONTENT REPRESENTATION BY INCREMENTAL NON-NEGATIVE MATRIX FACTORIZATION

Serhat S. Bucak *Bilge Gunsel*

Multimedia Signal Processing and Pattern Recognition Lab. Dept. of Electronics and Comm. Eng.
Istanbul Technical University 34469 Maslak Istanbul, Turkey

ABSTRACT

Nonnegative Matrix Factorization (NMF) is a powerful decomposition tool which has been used in several content representation applications recently. However, there are some difficulties in implementing NMF in on-line video applications. This paper introduces an incremental NMF (INMF) without deviating from conventional NMF's main objective function, which is minimizing the reconstruction error. The proposed algorithm is capable of modeling dynamic content of the video; thus controls contribution of the subsequent observations to the NMF representation properly. It is shown that the INMF preserves additive, parts-based representation capability of the NMF with a low computational load while offering dimension reduction. Experimental results are given to compare the reconstruction performances of the conventional and incremental NMF. In addition, video scene change detection and dynamic video content representation by INMF are investigated. Test results demonstrate that the INMF can be used as a powerful on-line factorization tool.

Index Terms— Non-negative matrix factorization, incremental algorithms, video content representation.

1. INTRODUCTION

High dimensional data usually contain an important amount of redundant components, which in fact may make the recognition process of vital components harder. Therefore finding suitable representations of data becomes extremely important in many data analysis tasks. Non-negative Matrix Factorization (NMF) [1, 2, 3] is one of these feature extraction/dimension reduction techniques, which approximately factorizes data into a form of multiplication of two non-negative matrices: matrix of basis vectors and encoding matrix. One of the major differences of NMF compared to other decomposition techniques is its constraint of non-negativity. This constraint, by allowing only additive combinations of the basis vectors, makes the NMF an intuitive, parts-based representation [1, 2, 3].

NMF has started to draw attentions with the work of Lee & Seung's [1] which alleged that NMF was successful in revealing parts-based features of data. Although Lee and Seung showed that non-negativity is a useful constraint for matrix factorization which can learn parts of the data, other works, after claiming that NMF cannot always guarantee parts-based representation in desired level, tried to increase sparseness in NMF in order to improve its localization capability [2, 3].

The NMF has started to find usage in many application areas recently [2]. With its success in revealing latent features in data and dimension reduction property, researchers started to use it more frequently in different fields such as face and object recognition, biomedical applications, document clustering, polyphonic music transcription, and color science. Even though the types of the applications using the NMF may differ, the way they employ it is quite the same.

The NMF, with its simple yet effective way to reduce dimension and extract intuitive features of interest, is a potential candidate for numerous video applications such as background modeling in surveillance type of video, video content analysis, etc. The conventional implementation of NMF is clearly not an on-line process since the NMF algorithm is executed once on the data matrix, constructed by the observed samples, and reaches to the final factorization at convergence. Therefore performing NMF continuously as each new frame arrives obviously will be computationally costly. Thus, a need for adaptations in NMF arises in order to make it available for on-line video applications. Being influenced by the work [4] which offers an incremental Principle Component Analysis (PCA), this paper introduces an incremental NMF (INMF) algorithm without deviating from the conventional NMF's main objective function, which is minimizing the reconstruction error.

The paper is organized as follows: In Section 2, after the necessary mathematical definitions are given, difficulties with the conventional NMF are discussed. In section 3, the incremental NMF is introduced. Finally, experimental results and conclusions are given in section 4.

2. THE CONVENTIONAL NMF

2.1. Mathematical Definitions

The aim of the NMF, with rank r being a pre-defined value, is to decompose the data matrix $\mathbf{V} \in \mathrm{R}^{n \times m}$ into two matrices; which are $\mathbf{W} \in \mathrm{R}^{nxr}$, also called as the mixing matrix, and $\mathbf{H} \in \mathrm{R}^{rxm}$, named as the encoding matrix [1, 2, 3].

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

The NMF is an iterative method which tries to find approximate factorizations as it is formulated in Eq.(1). Therefore the first step in factorization should be defining a cost function, so that with the intention of minimizing it, appropriate update rules to be used in each iteration could be defined for the elements of both $\mathbf{W}$ and $\mathbf{H}$.

Different cost functions have been defined in the literature, but because of its simplicity and effectiveness the squared reconstruction error given in Eq.(2) is used in this work.

$$F = \| \mathbf{V} - \mathbf{WH} \|^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(V_{ij} - (\mathbf{WH})_{ij} \right)^2 \tag{2}$$

The cost function F defined in Eq.(2) is a convex function of $\mathbf{W}$ and $\mathbf{H}$ separately, but not of both at the same time. Therefore, according to the gradient decent optimization, update rules which are performed alternatively for the elements of matrices $\mathbf{H}$ and $\mathbf{W}$ are given in Eq.(3) and Eq.(4), respectively [1, 2],

$$H_{aj}^{t+1} = H_{aj}^{t} \frac{\left(\mathbf{W}^{t^T}\mathbf{V}\right)_{aj}}{\left(\mathbf{W}^{t^T}\mathbf{W}^{t}\mathbf{H}^{t}\right)_{aj}} \tag{3}$$

$$W_{ia}^{t+1} = W_{ia}^{t} \frac{\left(\mathbf{V}\mathbf{H}^{t+1^T}\right)_{ia}}{\left(\mathbf{W}^{t}\mathbf{H}^{t+1}\mathbf{H}^{t+1^T}\right)_{ia}} \tag{4}$$

where t refers to the iteration number, T denotes the transpose, $a = 1,2,\ldots\ldots,r$; $i = 1,2,\ldots,n$, and $j = 1,2,\ldots,m$.

Note that Eq.(2) will be nonincreasing with respect to each element in $\mathbf{W}$ and $\mathbf{H}$ under the update rules given in Eq.(3) and Eq.(4) [1]. Initially elements of $\mathbf{W}$ and $\mathbf{H}$ are chosen as random nonnegative values.

2.2. Difficulties with the Conventional NMF

As a powerful technique which gives additive, parts-based representations of data as well as offering dimension reduction at the same time, the NMF with its previous success in various content analysis work [2, 3], is a prospect to be used in video applications, if it is described in an incremental form.

There are two striking points here about the conventional NMF that could cause problems in the video applications. First of all, as each new frame (sample) is taken, since the rank of the data matrix $\mathbf{V}$ is increased by one, the rank of the encoding matrix $\mathbf{H}$ is also increased, that yields a higher number of update operations per iteration. Since the number of operations increase nonlinearly, implementing the NMF repeatedly whenever a new frame arrives is not a practical solution at all.

Another problem with the conventional NMF is the need for original data matrix **V** in update operations, as it can be seen in Eq.(3) and Eq.(4). Storing the whole data matrix **V** throughout the whole process would require a significant amount of memory, especially when the sample size is high.

Therefore, a proper algorithm which is able to update the previous representations of video according to new-arrived frames without causing a heavy workload should be derived. To overcome these problems, we introduce an incremental NMF which is explained in the next section.

3. THE INCREMENTAL NMF

Since it wouldn't be very practical to carry out the conventional NMF for the whole video matrix as each new frame arrives, an incremental NMF representation which is appropriate to the on-line video applications is derived. The idea behind the incremental NMF (INMF) representation is altering the procedure in a way that prior representations get updated with each new frame.

In the incremental NMF, since each frame in **V** is reconstructed by the help of the corresponding column of the encoding matrix, a new frame automatically adds a new column to $\mathbf{H}$. Moreover, in each step, the mixing matrix $\mathbf{W}$ should be updated with the contribution of new frames. To achieve this, first of all, effect of the new frame (sample) on the cost function should be examined.

Let F defined in Eq.(2) be the cost function of m frames; thus is denoted as F_m. Consequently, $\mathbf{W}$ and $\mathbf{H}$ shown in Eq.(2) refer to $\mathbf{W}_m$ and $\mathbf{H}_m$, respectively. As the $(m+1)^{th}$ sample, **v,** arrives, a new component shown as f_{m+1}, which is used to formulize the reconstruction error of **v**, is added to the cost function as it is given in Eq.(5), where v_i refers the element of **v** and h_a denotes ath component of the new column of the encoding matrix. In Eq.(5) we introduce two new parameters: β and α. These parameters are crucial in controlling the algorithm's adaptability to dynamic content changes. α determines the influence of the last sample into the representation. β is introduced to limit the contribution of the old samples as the number of samples increases. When the condition $\beta \in (0,1)$ is satisfied, the effects of older frames decay continuously, allowing the new samples to participate more.

$$F_{m+1} = \beta F_m + \alpha f_{m+1} = \beta F_m + \alpha \sum_{i=1}^{n}\left(v_i - \sum_{a=1}^{r} W_{ia}h_a\right)^2 \tag{5}$$

In order to obtain a NMF representation for the new data matrix $\mathbf{V} \in \mathrm{R}^{n\times(m+1)}$, we need to minimize the F_{m+1} by minimizing both of its components. In other words, matrices $\mathbf{W}_m$ and $\mathbf{H}_m$ should be updated in a way that reconstruction errors for the old m frames and the new frame should be minimized that yields the best representation for all of the $(m+1)$ frames.

The cost function F_{m+1} defined in Eq.(5) is still a convex function of $\mathbf{W}_m$ and $\mathbf{H}_m$ separately, thus we can use the gradient descent algorithm to minimize the Eq.(5). Taking the derivative of Eq.(5) with respect to h_a, which corresponds to ath component of the new column **h** of the encoding matrix , gives Eq.(6).

$$\frac{\partial F_{m+1}}{\partial h_a} = -\alpha\left[\left(\mathbf{W}_m^T\mathbf{v}\right)_a - \left(\mathbf{W}_m^T\mathbf{W}_m\mathbf{h}\right)_a\right], \quad a = 1,2,..,r. \tag{6}$$

Thus, the update rule of gradient decent should be as in Eq.(7)

$$h_a^{t+1} = h_a^t - \eta_a \frac{\partial F_{m+1}}{\partial h_a^t} \tag{7}$$

where $\mathbf{h}^{t+1}$ denotes the **h** at iteration $(t+1)$ and the step size η_a is chosen as:

$$\eta_a = \frac{h_a^t}{\alpha\left(\mathbf{W}_m^{t^T}\mathbf{W}_m^t\mathbf{h}^t\right)_a} \tag{8}$$

Using Eq.(8) in Eq.(7) yields the update rule given in Eq.(9).

$$h_a^{t+1} = h_a^t \frac{\left(\mathbf{W}_m^{t^T}\mathbf{v}\right)_a}{\left(\mathbf{W}_m^{t^T}\mathbf{W}_m^t\mathbf{h}^t\right)_a} \tag{9}$$

Note that, the conventional NMF requires updating all the elements of $\mathbf{W}$ and $\mathbf{H}$, thus the number of operations increases nonlinearly as the number of frames increases. However, in our INMF derivation, there is no need to update the elements of encoding matrix for the previous frames, but only the components corresponding to the new frame are updated. Therefore at each iteration t, the number of elements to be updated and the number of

operations remain constant. Furthermore, the INMF updating rule does not require the original data matrix **V,** thus the INMF eliminates the need of storing whole video through the process.

On the other hand, derivative of Eq.(5) with respect to W_{ia} yields Eq.(10).

$$\frac{\partial F_{m+1}}{\partial W_{ia}} = -\beta\left(-\left(\mathbf{V}_m\mathbf{H}_m^T\right)_{ia} + \left(\mathbf{W}_m\mathbf{H}_m\mathbf{H}_m^T\right)_{ia}\right) + \alpha\left(-\left(\mathbf{v}\mathbf{h}^T\right)_{ia} + \left(\mathbf{W}_m\mathbf{h}\mathbf{h}^T\right)_{ia}\right) \quad (10)$$

With the step size λ_{ia} defined in Eq.(12), the gradient descent algorithm states:

$$W_{ia}^{t+1} = W_{ia}^t - \lambda_{ia}\frac{\partial F_{m+1}}{\partial W_{ia}^t} \quad (11)$$

$$\lambda_{ia} = \frac{W_{ia}^t}{\left(\mathbf{W}_m^t\left(\beta\mathbf{H}_m\mathbf{H}_m^T + \alpha\mathbf{h}^{t+1}\mathbf{h}^{t+1^T}\right)\right)_{ia}} \quad (12)$$

where $\mathbf{h}^{t+1}$ denotes the **h** at iteration (t+1).

Using Eq.(12) in Eq.(11) yields the update rule given in Eq.(13).

$$W_{ia}^{t+1} = W_{ia}^t\frac{\left(\beta\mathbf{V}_m\mathbf{H}_m^T + \alpha\mathbf{v}\mathbf{h}^{t+1^T}\right)_{ia}}{\left(\mathbf{W}_m^t\left(\beta\mathbf{H}_m\mathbf{H}_m^T\right) + \alpha\mathbf{h}^{t+1}\mathbf{h}^{t+1^T}\right)_{ia}} \quad (13)$$

Note that, since the matrices $\mathbf{v}_m$ and $\mathbf{H}_m$ remain the same at all iterations of the new sample's ((m+1)$^{\text{th}}$ sample) update procedure, storing the multiplications $\mathbf{V}_m\mathbf{H}_m^T$ and $\mathbf{H}_m\mathbf{H}_m^T$ instead of separate matrices reduces the computational complexity. Update iterations are repeated till convergence and the encoding matrix $\mathbf{W}_m$ is used as the initial state for running the algorithm when the (m+1)$^{\text{th}}$ sample is received. Note that, the influence of new and previous frames on the representation can be controlled by β and α.

4. EXPERIMENTAL RESULTS AND CONCLUSIONS

Experimental results are summarized in three subsections. In the first section, since the NMF aims minimizing the reconstruction error, a comparison between the conventional and incremental NMF with respect to their reconstruction performances is done. In the second section, background modeling is given as an example in order to show the incremental NMF's effectiveness in adapting to content changes. Finally, performance in video scene change detection is evaluated to demonstrate a potential use of INMF.

4.1. Reconstruction Performance of the INMF

The conventional NMF tries to minimize the reconstruction error given in Eq.(2). Therefore, the primary comparison between the conventional NMF and INMF has been made in terms of their reconstruction performances. In order to observe reconstruction error with respect to frame number, both the NMF and INMF are applied on a 150 frames surveillance type of video clip where the scene does not experience big changes. First, the NMF representation with r=5 was obtained after initial **W** and **H** were specified as random nonnegative values and the convergence has been reached at t= 1000 iterations. Then, the INMF was applied on the same 150 frames with rank fixed to 2. The same initial **W** and **H** matrices were used to initiate INMF iterations. Convergence has been reached at about t=15 iterations per frame. Fig.1 illustrates reconstruction error f_m for each frame. To justify the effect of the weighting parameters, the INMF is applied at 2 different pairs of parameter, (α, β) = (0.2, 0.8) and (0.8, 0.2). As it is expected, reconstruction error obtained by the introduced INMF is less than the NMF. Especially for $\alpha = 0.8$, the INMF outperforms the rest at all frames. This is because its capability of adapting a new frame into the factorization in an on-line manner. However, the NMF reconstructs each individual frame by using the basis matrix **W** obtained via the batch processing of 150 frames.

4.2. Dynamic Content Representation by INMF

Background modeling in surveillance type of video is a good example to judge the dynamic content representation performance of a statistical method. This requires representation of the background scene by using a small number of background frames and then updating this representation in such a way that dynamic content changes influence the representation appropriately. These changes include entrance / leaving of an object into / from the scene and detection of changes in object motions.

In the literature a number of work has been performed for the representation of background by using Incremental Principal Component Analysis (PCA) [4]. Influenced by this work, our recent work in [5] tackles with this problem via the introduced INMF in more detail. Conventionally, a quantitative measure of the dynamic content representation capability is the reconstruction error between the original frame and its reconstruction based on the background representation of the scene. It is expected to obtain a significant increase in the error if there is a change in the scene while the error converges to zero for the background frames. Of course this is true under the assumption that the representation of background is satisfactory.

Performance of INMF has been evaluated on a video clip from the PET2001 video surveillance dataset [4]. At frame 1930, the scene includes an initially stationary car which starts to move at frame 1990. In addition to this car, there are also 2 walking men in the scene. After modeling the background initially, the foreground objects are tried to be separated from the dynamic background by using the representations obtained by NMF and INMF. Fig.2, which corresponds to Frame 2061, is given to compare the performances. Note that, as a result of updating the background model dynamically, the INMF is capable of removing the stationary car from the background as soon as it starts moving and treating it as a foreground object, whereas NMF fails to do so. While the same conclusion is valid for the walking men, the flu sight of the background in the picture also proves that NMF fails to represent it as successful as INMF. The INMF parameters for this test were $\beta = 0.8$, $\alpha = 0.2$ and r=2 for both. 10 frames were used for representation of the background for both of the methods.

4.3. Video Scene Change Detection by INMF

Aim of the video scene change detection problem is to be able to detect the frames where scene changes take place. These scene changes can be classified as "cuts" and "gradual changes." Clean cuts are sudden changes between the scenes. In contrast, gradual changes, i.e., fade in/out, dissolve, wipe, etc., are generally longer

and can be defined as continuous transitions between two different video scenes. Obviously, detection of the gradual changes is a more difficult task. Another difficulty is avoiding false alarms which are likely to be caused by camera and object movements or lighting variations throughout a scene.

In the literature a number of video scene change algorithms have been reported [6]. In this subsection a potential use of INMF in video scene change detection has been evaluated. The idea behind is that if the INMF accurately represents the scene content and the dynamic changes, a significant increase in the reconstruction error should be detected at the scene change frames. Thus, determination of a video scene change is done by examining the changes on the reconstruction errors of the successive frames.

Tests are carried out on over 130000 (%24 of all the transitions were gradual) frames from the video clips recorded in TRECVID database [7]. In order to reduce computational complexity, the INMF has been performed on the DC images of each frame when $\beta=1$ and $\alpha=1$. Table 1 reports the video scene change performance of the INMF in terms of "precision" and "recall" [7]. Precision is defined as the ratio of correct matches to the total number of transitions reported. On the other hand, recall is the number of correct matches divided by the total number of actual transitions in the video sequence. Hence, precision gives clue about the system's false positive performance whereas recall is related to false negative ratio. As it is shown in Table 1, the number of false alarms is small, detection rates for both gradual transitions and cuts are high, imposing that the INMF is a promising tool in content analysis. However note that, the results are obtained by a supervised control mechanism and more tests with an increased number of gradual changes should be performed in order to estimate the performance in more detail.

It is concluded that the introduced incremental non-negative matrix factorization, with its ability to adapt the conventional NMF's useful features to its incremental nature, is an efficient tool for modelling dynamic content in video applications. Besides making new tests on scene change detection, currently we are working on derivation of sparse INMF which could be more beneficial to have more localized, parts-based representations to increase robustness to lighting variations and motion.

REFERENCES

[1] D.D. Lee and H.S. Seung, "Learning the Parts of Objects by Nonnegative Matrix Factorization," *Nature*, vol. 401, pp. 788-791, 1999.

[2] A. Pascual-Montano, J.M Carazo, K. Kochi, D. Lehmann, and R.D. Pascual-Marqui, "Nonsmooth Nonnegative Matrix Factorization," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 42, pp. 403-415, 2006.

[3] P.O. Hoyer, "Non-negative Matrix Factorization with Sparseness Constraints," *Jornal of machine Learning Research*, vol. 5, pp.1457-1469, 2004.

[4] Y. Li, J. Xu, L. and Morphett, and R. Jacobs, "An Integrated Algorithm of Incremental and Robust PCA," *Proc. of IEEE Int. Conference on Image Processing*, Barcelona, Spain, 2003.

[5] S.S, Bucak, B. Gunsel, O. Gursoy, "Incremental nonnegative matrix factorization for dynamic background modeling," *Proc. of ICEIS Int. Workshop on Pattern Recognition in Information Systems*, Funchal, Portugal, June 2007.

[6] B. Gunsel, A. M. Tekalp, and P. van Beek, "Content-based access to video objects: Representation, temporal segmentation, and feature extraction," Signal Processing, vol.66, issue 4, pp.261-280, 1998.

[7] TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/

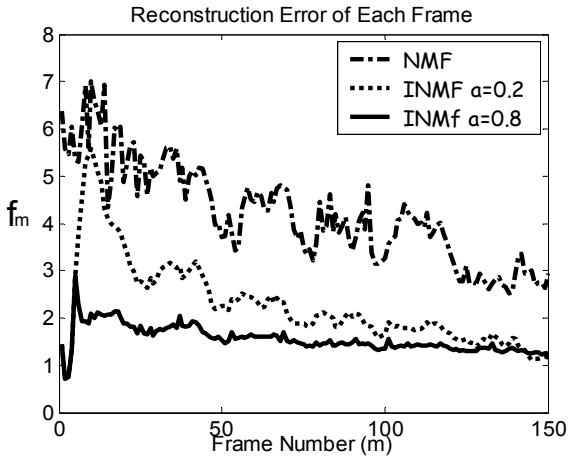

Fig. 1. Distribution of the reconstruction error wrt frame number.

Table 1. Scene change detection performance of the INMF.

	Cuts	**Graduals**	**Total**
True Positives #	590	174	764
False Negatives #	42	28	70
Recall	0.93	0.86	0.92
False Positives #			120
Precision			0.86

Fig. 2. a) Original frame no:2061. b) Incremental NMF (r=2, α=0.2, β=0.8). c) Conventional NMF (r=2).