Concept Detection in Images Using SVD Features and Multi-Granularity Partitioning and Classification

Kamran Farajzadeh
Department of IT management, Islamic Azad University, Science and Research Branch, Tehran, Iran
k.farajzadeh@iau-tnb.ac.ir
Esmail Zarezadeh
Department of Electrical Engineering, Amir Kabir University, Tehran, Iran
zarezadeh@aut.ac.ir
Jafar Mansouri*
Department of Electrical Engineering, Ferdowsi University of Mashhad, Mashhad, Iran
jafar.mansouri@gmail.com

Received: 24/Oct/2016 Revised: 08/Aug/2017 Accepted: 20/Aug/2017

Abstract

New visual and static features, namely, right singular feature vector, left singular feature vector and singular value feature vector are proposed for the semantic concept detection in images. These features are derived by applying singular value decomposition (SVD) "*directly*" to the "*raw"* images. In SVD features edge, color and texture information is integrated simultaneously and is sorted based on their importance for the concept detection. Feature extraction is performed in a multi-granularity partitioning manner. In contrast to the existing systems, classification is carried out for each grid partition of each granularity separately. This separates the effect of classifications on partitions with and without the target concept on each other. Since SVD features have high dimensionality, classification is carried out with K-nearest neighbor (K-NN) algorithm that utilizes a new and "*stable*" distance function, namely, multiplicative distance. Experimental results on PASCAL VOC and TRECVID datasets show the effectiveness of the proposed SVD features and multi-granularity partitioning and classification method.

Keywords: High-Dimensional Data; Multi-Granularity Partitioning and Classification; Multiplicative Distance; Semantic Concept Detection; Static Visual Features; SVD.

1. Introduction

Semantic concept detection in images is the process of deriving meaningful terms that describe image contents. It is also referred to as image annotation [1] or indexing [2]. Semantic concept detection has been an active research topic in the recent years due to its potentially large impact on the image understanding, summarization, search, and filtering. It is essentially a classification task for determining the presence of the given semantic concepts in an image. The semantic concepts cover a wide range of topics such as those related to objects (e.g., car, airplane), indoor/outdoor scenes or locations (e.g., meeting, desert), and genre (e.g., weather, sports).

The success of a concept detection scheme strongly relies on the effectiveness of the low-level features in the content representation. Many systems have used global features mostly specified by MPEG-7 Visual part [3]. Global features include color or edge histograms, grid-based color moment and wavelet texture, etc. Other widely-used features are local features, like scale invariant feature transform (SIFT). Local features represent an image by the histogram of local patches based on a visual vocabulary of visual words [4]. An image is decomposed to a set of visual words derived after clustering or segmentation of the input image. However, local and global features have their own weaknesses. Global features do not contain local structure, and local features do not represent statistics about the overall distribution of texture or edge information. Moreover, most local features, like SIFT, are dependent on the size of the vocabulary of visual words; and the size of the vocabulary is also dependent on the type of images. For different types of data, the suitable size of the vocabulary changes.

This paper has the following contributions: New static visual features, namely, singular value feature vector, right singular feature vector, and left singular feature vector, are proposed. These features are derived by applying SVD "*directly*" to the "*raw*" images. In SVD features, different information like color, edge and texture is incorporated into an integrated framework and this information is sorted in accordance with their importance for detecting concepts. Moreover, feature extraction and classification is performed in a multi-granularity scheme, which is approximately similar to what is done by a human for detecting a concept. Classification is carried out for each grid partition of each granularity separately. Furthermore, feature vectors usually have high dimensionality even after dimension reduction. Recently, it has been shown that when dimensionality of data (feature vectors) is high, many distance functions (Minkowski distances, cosine similarity, etc.) become unstable [5][6][7][8]; i.e. distances

* Corresponding Author

of all data to a given query point become the same in the high-dimensional space. This phenomenon leads to the performance degradation of the classification algorithms that use these distance functions. To solve the instability problem, a new distance function, namely, multiplicative distance [8], is used that its stability in the high-dimensional space has been proved.

The rest of the paper is organized as follows. Section 2 gives an overview on the related works. Section 3 states characteristics of SVD features. This Section also introduces multiplicative distance. In Section 4, the proposed concept detection method is demonstrated in detail. The experimental results are reported in Section 5. Finally, some conclusions are drawn in Section 6.

2. Related Works

Generally speaking, two types of static visual features are often used: global and local. While global features are statistics about the overall distribution of color, edge or texture information, local features describe the local structures in an image. In the following, we mention some of these features in the previous works. Many papers like [9][10] have used global features, such as edge direction histogram, Gabor texture, color moment, color histogram, canny edge, etc., for describing images. Most of these features are defined by MPEG-7 Visual part. These features either are concatenated to form a single feature vector for the classification [9] or are used separately for the classification [10]. In both of these cases, the relationships between the features are not usually taken into account. In [11] local binary pattern (LBP) has been used for image feature description. In [12] an image descriptor based on the orientation of contrasts is proposed. The contrast and width of canny edge features are computed by a simple scheme. Inspired by the histogram of oriented gradients method, an image representation is proposed based on a histogram of contrasts.

Some further research works fall in the category of the local features. An image has local interest points or keypoints defined as salient patches that contain rich local information about the image. Keypoints are usually around the corners and edges of image objects, such as the edges of the map, people’s faces, etc. The most popular keypoint-based representation is bag-of-visual-words (BoW). In BoW, a visual vocabulary is generated through grouping similar keypoints into a large number of clusters and treating each cluster as a visual word. An image can be represented as a histogram of visual words. The performance of BoW features in semantic concept detection is subject to various representation choices [4].

Lazebnik et al. [13] have exploited the spatial location of keypoints and proposed a spatial pyramid matching (SPM) method, in which an image is first divided into multi-level equal-sized grids and each grid is described by a separate BoW using SIFT descriptor. Then, the BoWs from image grids at each level are concatenated to form the final representation. Fisher vector is another type of the image representation that has been used in [14]. The Fisher vector can be seen as an extension of the BoV. Both of them are based on the visual vocabulary built on the low-level features like SIFT descriptor. If a Gaussian mixture model is used to model the visual vocabulary, the gradient of the log likelihood can be computed with respect to the parameters of the model to represent the image. Sparse coding is a feature encoding method that has been widely-used in recent years [15]. The aim of sparse coding is to represent input vectors as a linear combination of a small number of basis vectors (dictionary).

In [16] a method called non-negativity and locality constrained Laplacian sparse coding is proposed. Firstly, non-negative matrix factorization is used in the Laplacian sparse coding, which is applied to constrain the negativity of both codebook and code coefficient. Secondly, K-nearest neighboring codewords for local features are used because locality is more important than sparseness. Finally, non-negativity and locality constrained operators are utilized to obtain a novel sparse coding for local features. Convolutional neural network (CNN) has been used in [17]. It consists of multiple convolutional layers of small neuron collections followed by fully connected layers. These layers form multi-stage feature extractors, which higher layers generate more abstract features from lower ones. The input to the CNN is raw image pixels such as an RGB vector, which is forwarded through all feature extractor layers to generate a feature vector that is a high-level abstraction of the input data.

Some papers have used combination of global and local features. Jiang et al. [4] have used local BoW feature and two types of global features, color moment and wavelet texture, which have been obtained from the whole image. The local and global features have been combined and classification has been performed for the whole image, neither for each partition nor for each granularity.

3. Preliminaries

3.1 Motivations for Using SVD Elements as the Low-Level Feature

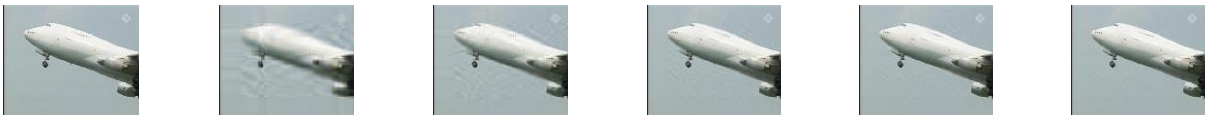
SVD elements can be useful for expressing edges, textures and colors/luminance in images. For describing this matter, notice that an image, an $m \times n$ matrix A, can be interpreted as the ensemble of the basis images as follows:

$$A_z = \sum_{i=1}^{z} \sigma_i u_i v_i^T \tag{1}$$

Where $z (z \leq p, p = \min(m, n))$ is the number of u_i (left singular vector) and v_i (right singular vector) pairs used. Each $u_i v_i^T$ specifies a layer of the image geometry, whereas the singular value σ_i is the weight assigned to this layer and specifies the luminance of that image layer [18][19][20]. The first few singular vector pairs account for the major image structure, whereas the subsequent u_i and v_i pairs account for the finer details in the image.

Larger singular values indicate more energy in the image. The singular values also denote the activity level in the image. A high activity level represents roughness or strong textures and edges. Similarly, a low activity level corresponds to smoothness or weak textures and edges [18][20]. These show that SVD can be used for representing different regions of an image.

If the first few singular values have predominant magnitude, after projecting along the first few singular vector pairs the reminder scatters are small and ignorable. This point is demonstrated through an example shown in Figure 1, where the image size is 352×288 and thus *p*=288. It can be observed that the first 30 basis images (*z*=30) [i.e., i = 1 to 30 in (1)] capture the major image structures and luminance/colors, and the subsequent basis images signify the finer details in the image. Furthermore, it is observed that with some few singular values and their respective left and right singular vectors, an image can be reconstructed nearly similar to the original image.



(a) original (b) z =10 (c) z =20 (d) z =30 (e) z = 40 (f) z = 50

Fig. 1. A_z as defined by (1) for different values of Z, (a) original image ($Z = 288$), (b) Z= 10, (c) Z = 20, (d) Z = 30, (e) Z =40, (f) Z =50.

In SVD, color/ luminance, texture, and edge information is sorted according to their significance. SVD integrates this information simultaneously and takes into account the relation between them. Hence, all the edge, color and texture information is encoded into a single representation. In addition, there is no redundant information in SVD since left and right singular vectors are orthonormal. Furthermore, SVD takes into account human visual perception [20].

Moreover, singular values are stable. The stability of singular values indicates that when there is a little disturbance in the image, singular values do not change considerably [18]. Therefore, singular value features can be effective to encounter noise, small clutters and small changes in the image. Additionally, singular values are useful when the image has transposition, rotation and translation; since a) a matrix A of an image and its transpose, A^T, have the same non-zero singular values [18]; b) if R is a unitary and rotating matrix, the singular values of *RA* (rotated matrix) are the same as those of A [19]; c) the original image A and its rows or columns interchanged image have the same singular values [19]. These abilities motivate us to use SVD elements as the low-level feature for the concept detection.

3.2 Multiplicative Distance

Under some conditions on the data distribution, distances between data and query points in the high-dimensional space are meaningless or unstable [5][6][7][8]. This means that distances of all data from a given query point become the same for a wide variety of data distributions and distance functions, when dimensionality increases toward infinity. Minkowski and fractional norm distances [8], cosine similarity for the i.i.d. (independent and identically distributed) data [21] are some examples. In such cases, the concept of proximity and similarity is not meaningful because of the poor discrimination between the nearest and furthest neighbors. This instability can greatly affect many applications like classification, and can result in the performance degradation. In [6] and [7] authors have stated that the sufficient and necessary condition for the stability of a distance function. In [8] a new distance function, namely, multiplicative distance, has been introduced that its stability has been proved. The multiplicative distance function can be used in the low-dimensional space. The definition of this distance function is as follows.

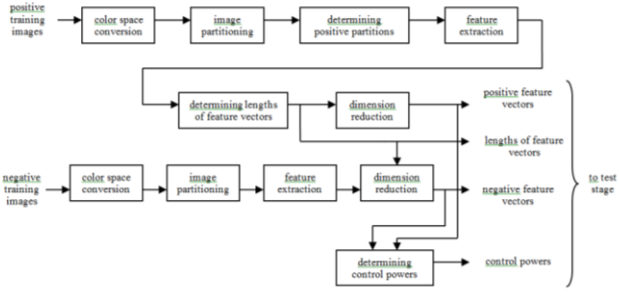
Definition: Let $X = (x_1, x_2, \dots, x_m)$ be an m-dimensional random vector with $x_k \sim F_k$ (F_k is the distribution of the random variable x_k), $k = 1, \dots, m$ and $Q = (q_1, q_2, \dots, q_m)$ be the query point with $q_k \sim \widetilde{F_k}$. Set $z_k = 1 + |x_k - q_k|$. The general form of the multiplicative distance of X from Q is defined as:

$$\mathrm{MD}(X, Q) = \left(\prod_{k=1}^{m} z_k^{c_k} \right) - 1 \qquad (2)$$

where c_k is named "control power", which controls the effect of each z_k on the distance. z_k^{ck} is defined as distance component. If $\forall k \colon c_k = c$, each dimension has equal effect on the distance. In the simple form of the multiplicative distance we have $\forall k \colon c_k = 1$.

4. Proposed System

Figure 2 shows the block diagram of the proposed concept detection for each case of multi-granularity partitioning, which includes training and test stages.



(a)

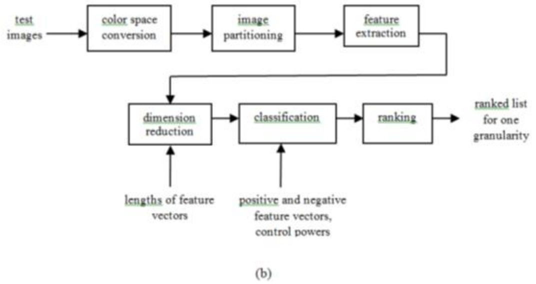


Fig. 2. Block diagram of the proposed concept detection for one case of multi-granularity partitioning, (a) training stage, (b) test stage.

4.1 Training Stage

First, positive images (i.e. images containing the target concept) are partitioned into different granularities. Usually, six cases of granularities, i.e. 1×1 grid (1 partition), horizontal 2×1 grid (2 partitions), vertical 1×2 grid (2 partitions), 2×2 grid (4 partitions), 3×3 grid (9 partitions) and 4×4 grid (16 partitions) are used for feature extraction for most of concepts. Since some concepts are depicted by the holistic representation of an entire image rather than a region (e.g., office), very small granularities (like 4×4 grid) are not used for these concepts.

This kind of partitioning is approximately similar to what is done by the human for detecting concepts. When a human looks for a concept in an image, in accordance with the concept, first, he/she looks at the whole image and if it is necessary, the view range is getting smaller in order to detect the concept. So, the human hierarchically narrows down the looking range until detecting the target [22][23]. However, the human perception system is very complicated and uses extensive rules. In the proposed method, the similar work is done but in a simple way. The positive images are divided into equal grid partitions at different granularities so that they are perceptible for the human for detecting concept. Additionally, very small partitions are not used since they have no perceptual meaning.

Different types of partitions based on multi-granularity partitioning are obtained. Then, partitions that contain the concept (positive partitions) are detected manually and are kept. As stated in Section 3, SVD elements can be used as the low-level features. Therefore, SVD is applied to each positive partition of the image. SVD features are extracted in three cases: from "*raw*" color images with RGB and HV color spaces and "*raw*" gray scale space. The right singular vectors of SVD are concatenated to form the right singular "feature" vector. The left singular vectors of SVD are also concatenated to form the left singular "feature" vector. Singular values are also put in the singular value "feature" vector. Therefore, in the case of experiments with the gray scale space, there are 3 feature vectors for each partition; and for the case of experiments with the HSV and RGB color space, there are 9 feature vectors (3 feature vectors for each component of the color space).

Notice that SVD features are obtained "*directly*" from "*raw images*" of partitions. Our SVD features are different from features in methods that use techniques like PCA or LSA. In those methods, low-level features (like SIFT features) are extracted. Then, SVD in techniques like PCA or LSA is applied to the "*extracted low-level features*" (not on the "*raw images*") to produce the final processed features.

SVD features are obtained from the positive partitions. However, feature vectors usually have very high dimensionality. High-dimensional data requires large storage space and more computation and to reduce the computational and storage cost, dimension reduction is necessary. With the dimension reduction, some data that do not help for detection are removed and performance can be improved. In addition, the dimension reduction is useful for the noise cleaning. An interesting property of SVD is that information is sorted based on its importance in the descending order. Small or zero singular values indicate that their respective right and left singular vector pairs (in Equation (1)) have less or no significance. Furthermore, if the first few singular values have a predominant magnitude, after projecting along the first few singular vector pairs, the remainder scatters can be ignored. Therefore, with removing less or insignificant singular values and vectors, the dimensionality is reduced. For this task in the training stage for each positive partition (of each case of granularities) that contains the target concept, its energy is calculated from the below formula:

$$E = \sum_{i=1}^{p} \sigma_i^2 \qquad (3)$$

where $p = \min(m, n)$, m and n represent the size of the partition. Then, the first index of the singular values that satisfies the following equation is obtained:

$$E_{th} = th \times E \leq \sum_{i=1}^{index} \sigma_i^2 \qquad (4)$$

where th is a threshold. E_{th} is the energy of the reconstructed image (partition). th is selected so that after the dimension reduction, the reconstructed image has good perceptual quality and no considerable distortion. For simplicity, th is selected the same for all concepts (for all granularities and also all components of color images). For each case of partitioning these indices are calculated separately for all partitions containing the target concept, and the final index for each granularity, which is used in test stage, is the average of these indices. Thus, we have different indices for different granularities (e.g., 6 final indices for each kind of feature vector in the case of 6 granularities with gray scale images). For HSV and RGB cases, for each color component of the images this process is performed separately. Therefore, usually different indices related to th are yielded for three components of the color images. These final indices determine lengths of feature vectors for the concept. With removing the singular values after these indices and their respective elements in the left and right singular feature vectors (i.e. their respective left and right singular vectors in Equation (1)), the dimensionality of feature vectors are reduced and final positive feature vectors are yielded. The reduced-dimension feature vectors of the proposed system have different lengths for different concepts.

Based on Figure 2, for the negative images (images without the target concept), multi-granularity partitioning is performed. Then, SVD features are obtained from partitions. Using lengths of feature vectors (from obtained indices), the dimension reduction is carried out on these features and the final negative feature vectors are attained. Furthermore, since for the classification in the test phase the multiplicative distance is used, the parameter *c* (control power) for the multiplicative distance is selected such that the overflow does not happen. For this purpose, for each granularity and for each type of feature vector for each concept, the pairwise multiplicative distances of all feature vectors (positive and negative) for different values of c are calculated. The largest value of the non-positive integer powers of 10, i.e. 1, 0.1, 0.01, ..., for which overflow does not happen, is selected as the control power. The final positive and negative feature vectors, lengths of feature vectors and control powers are used in the test stage.

4.2 Test Stage

For each test image, multi-granularity partitioning is done. Then, SVD is applied to the partitions of raw images and SVD features are obtained from each partition. The dimension reduction is performed on the feature vectors of each partition using the lengths of feature vectors (final indices from the training stage). However, even after dimension reduction the feature vectors usually have high dimensionality. As stated in Section 3.2, conventional distance functions in the literature become unstable for the high-dimensional data. So, using these distances in a classifier can result in the performance degradation. Multiplicative distance has been introduced as a stable distance function [8] and we use this distance in our work. Notice that a classifier must be used that this distance function is applicable to it. For this reason, classification is carried out with the well-known K-NN algorithm with the stable multiplicative distance function.

If we consider features of all partitions of an image together for the classification, some partitions may not have the target concept. Therefore, wrong detection can occur due to considering features of the partitions with and without the target concept together. Thus, in the proposed method classification is carried out for each grid partition of different granularities individually. Furthermore, classification is performed for each of 3 or 9 kinds (based on gray scale or color images) of feature vectors of each partition separately.

It should be noted that our multi-granularity partitioning and classification system is different from the spatial pyramid matching approach, commonly used in BoW representation like in [13]. In [13] an image is partitioned into different granularities. For each granularity, BoW features are obtained. Finer granularities get higher weights. Then, features of all granularities are concatenated and classification is performed for the whole image not for each partition. In our work in contrast to [13], features of different granularities are not concatenated and detection is performed for each grid partition of each granularity separately. Furthermore, different granularities have the same importance in detection. This is because finer granularities do not necessarily represent concepts better. Moreover, for some concepts we do not use small granularities as stated before.

If each feature vector has n dimensions and number of the training feature vectors is m, and $X_i = (x_{i,1}x_{i,2}, \dots, x_{i,n}$ $i = 1, \dots, m$ are the training feature vectors, and $Y = (y_1, \dots, y_m)$ is a feature vector of the test sample, the classification is as follows:

$$label \quad Y = label \quad \arg \quad \min_i \prod_{j=1}^{n} \left(|x_{i,j} - y_j| + 1 \right)^c$$

$$f(Y) = \min_i \prod_{j=1}^{n} \left(|x_{i,j} - y_j| + 1 \right)^c \qquad (5)$$

where f is the distance output of the classification. *label* of Y is positive if X_i that has the minimum distance is from the positive training feature vectors; otherwise *label* is negative. C is the control power of the multiplicative distance. For all 9 (for RGB and HSV images) or 3 (for gray scale images) feature vectors of a partition, the classification is performed separately. For a partition, if at least one of the classifications on its feature vectors gives positive answer for the target concept, that partition is annotated with the positive label. For an image, if at least one of its partitions is positive, that image is annotated with positive label for that granularity.

For each concept, following stages are performed for "each granularity":

a. For each positive image the partition(s), namely, the best partition (s), with the maximum number of positive labels of the classifications is (are) kept as the representative of that image and other partitions are eliminated.
b. The best partitions of positive images are divided into 3 or 9 groups in a descending order in accordance with the number of their positive labels. For example, for the case of gray scale images, there are 3 groups: the first group is related to the best partitions that have 3 positive labels; the second group contains the best partitions with 2 positive labels; and the last group is related to the best partitions with 1 positive label.
c. Among each group of best partitions, for each "kind" of feature vector, feature vectors (with positive or negative label) are ranked based on their labels and distances, *f*. First, positive feature vectors are ranked in an "ascending" order according to their distances. Next, negative feature vectors are ranked in a "descending" order based on their distances. The score of each best partition is the summation of ranks of all its feature vectors.
d. If an image has some best partitions (i.e. with the same number of positive labels), the partition with the best score is kept and other partitions of that image are eliminated.
e. The best partitions are ranked according to their group and score, to form the ranked list of positive

images for that granularity. Note that each best partition is related to one image.

Next, the ranked lists of all granularities are aggregated with simple fusion to form the final ranked list of positive images for that concept.

5. Experimental Results

In this Section, the performance of the proposed method is evaluated on two well-known PASCAL VOC and TRECVID datasets. The first dataset is the widely-used PASCAL VOC 2007 [24], which consists of 8 concepts and includes 9963 images divided into a predefined training and test set of 5011 and 4952 images, respectively. Figure 3 shows the example images for 8 concepts in PASCAL VOC 2007.

Fig. 3. Exemplary images for the evaluated concepts in the experiments in PASCAL VOC 2007 dataset.

The images are resized to 352×288. For each semantic concept, average precision (AP) is used for the performance evaluation. AP is the average of precisions computed at the point of each of the relevant (by the ground truth annotation) images for considering the order in the ranked list of images [25]. To evaluate the overall performance, we use mean average precision (MAP) which is the mean value of the APs over all concepts.

The second dataset is the keyframes of TRECVID 2007 (TV07) [26]. The national institute of standards and technology (NIST) has established "semantic indexing" as a task in TREC video retrieval evaluation (TRECVID) [26], which aims to provide a benchmark for evaluating video concept detection technologies. The shot and sub-shot detection and extraction of keyframes have been performed by TRECVID. The training and test datasets consist of 21532 and 22084 keyframes, respectively. The keyframes are in CIF (352×288) format. There are 20 semantic concepts evaluated in TV07. Figure 4 shows the example keyframes for all 20 concepts evaluated by TRECVID. The reason for selecting TV07 is that all concepts are detectable by visual features.

Fig. 4. Exemplary keyframes for the evaluated concepts in the experiments in TRECVID 2007 dataset.

For each concept, inferred AP average precision (infAP) is used for the performance evaluation. InfAP is designed for partially labeled datasets (like TV07 test set). The infAP is an approximation of the AP and can save significant judging effort during the annotation of ground truth for large test dataset [25]. For each concept, infAP is computed based on the returned rank list and the ground truth provided by TRECVID. Notice that for the semantic concept detection in TRECVID, since the final annotation has been carried out for shots, if one shot has some positive sub-shots, just the sub-shot whose keyframe has the better rank is kept and other sub-shots are removed. Note that each keyframe is related to one sub-shot. Following the TRECVID evaluation, the infAP is computed over the top 2000 ranked shots according to the outputs of the proposed system. The mean infAP (MinfAP), which is the mean value of the infAPs over all concepts, is used for evaluating the overall performance.

First, we compare the proposed multi-granularity partitioning and classification method, which is referred to as MGPC, with the spatial pyramid matching (SPM) method [13] that is the most well-known partitioning and classification scheme and similar to our method. In SPM an image is first divided into multi-granularity equal-sized partitions and each partition is described by a separate BoW using SIFT descriptor. Then, the BoWs from image partitions at different granularities are concatenated with weights proportional to the level of the granularities to form the final representation for that image. Notice that bigger granularities (higher levels) have higher weights. For having a fair comparison, for both of the MGPC and SPM, the SIFT descriptor is selected as the low-level feature and two classifiers, the K-NN algorithm with the multiplicative distance and SVM with RBF (radial basis function) kernel are used for both methods. Notice that validation experiments have been performed for selecting K in K-NN (for K=1,3,5,7,9) and parameters C and gamma in SVM (C=$2^{-5},\dots,2^{4}$ and $\gamma = 10^{-7},\dots,10^{2}$). For brevity of the paper, just the final results are reported. Based on the experiments, for K-NN, K=5 and for SVM, C=4 and $\gamma = 0.1$ are obtained. The selection of control power in the multiplicative distance has been illustrated before in the training stage.

It is important to note that partitioning is continued until partitions are perceptible for human detection of a specific concept. For PASCAL VOC dataset, all 6 granularities are used for all concepts. But for TV07 dataset, some cases of granularities are not used for some concepts. For "people_marching" and "meeting", 4×4 and 3×3 grid cases and for "office", 4×4, 3×3 and 2×2 grid cases are not used. This is because it seems that partitions of these cases cannot represent these concepts individually. For example, one partition in 3×3 grid case usually cannot represent the concept of "office" in an image. Table 1 presents the definition of fusions, i.e. number of granularities used. For example, for MGPC method, fusion5 means that 5 granularities are used for partitioning. Notice that for the SPM method, definition of fusions in Table 1 means the "levels of spatial pyramids" used in [13]. Figures 5 and 6 show the performance of the proposed method and SPM method for PASCAL VOC and TRECVID datasets for different fusions.

From Figures 5 and 6 it is observed that the proposed method has the superior performance over the SPM method. The reason is that in SPM, features of partitions of different granularities are combined to form the final feature vector. Then, just one classification is carried out on the final feature vector. Thus, features of the partitions “without” the target concept can affect features of the partitions “with” the target concept and this can lead to the wrong result in the classification. However, in the proposed method the above problem does not occur. The reason is that the classification is perform for each partition of different granularities separately and features of one partition does not affect features of the other ones. If one partition of a granularity contains the target concept, the result of classification on its features will be positive. Therefore, that image is labeled as positive (with the target concept) and that image is ranked in accordance with the best value of all the positive classifications of different granularities.

Furthermore, in our method, different granularities have the same worth in detection but in SPM bigger granularities have higher weights (for their features) which is not true since bigger granularities do not necessarily represent the concept better than the smaller granularities. In addition, it is observed that using “very” small granularities leads to the decrease of the performance. Furthermore, for both of the proposed and SPM methods, the K-NN classifier gives the better results than the SVM classifier. The reason is that the K-NN uses the multiplicative distance which is stable for the high-dimensional space (even for SIFT features). However, the SVM uses the radial basis function with “Euclidean distance” that its instability leads to the performance degradation.

Of course, the proposed method needs the manual detection of positive partitions in the training stage, which makes it labor-expensive for huge training datasets. If it is necessary, for solving this problem we can use semi-supervised algorithms or sampling techniques. On the other hand, the SPM is an unsupervised method and does not require human labor.

Table 1. Notations for different cases of fusions of different granularities.

Notation	**Cases of fusion**
fusion1	1×1 + 1×2
fusion2	1×1 + 2×1
fusion3	1×1 + 1×2 + 2×1
fusion4	1×1 + 1×2 + 2×1 + 2×2
fusion5	1×1 + 1×2 + 2×1 + 2×2 + 3×3
fusion6	1×1 + 1×2 + 2×1 + 2×2 + 3×3 + 4×4

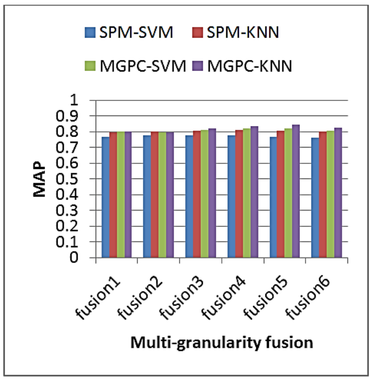


Fig. 5. MAP for different kinds of multi-granularity fusions for the proposed and SPM methods with SVM and K-NN classifiers for PASCAL VOC dataset.

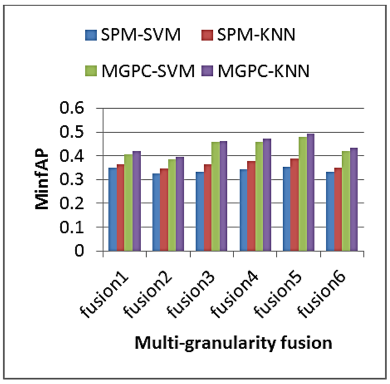


Fig. 6. MinfAP for different kinds of multi-granularity fusions for the proposed and SPM methods with SVM and K-NN classifiers for TRECVID dataset.

Now, we consider SVD features for different color spaces. For notational simplicity, the proposed method, which uses SVD features with multi-granularity partitioning and classification scheme, is referred to as SVDMGPC. Therefore, SVDMGPC-gray, SVDMGPC-RGB and SVDMGPC-HSV refer to the proposed method applied to images with 3 cases of gray-scale, RGB and HSV color spaces, respectively. In the experiments, the parameter *th* for the dimension reduction is set as 0.998 for all concept and all components of the color space. This value is obtained manually by subjective analysis for some values of *th*. However, it is possible that value of *th* is chosen adaptively for each concept using the objective quality metrics.

Figures 7 and 8 show the performance of the proposed method for different granularities individually for PASCAL VOC and TRECVID datasets. As shown in Figures 7 and 8, for both of datasets the vertical 1×2 grid has the best result among the granularities and 4×4 grid has the worst result. Moreover, the performance decreases for small granularities, i.e. 3×3 and 4×4 grids. This states that small granularities are not good choices for partitioning since small partitions may be not able to represent the target concept especially for non-object concepts (like in TRECVID dataset).

Figures 9 and 10 show the performance of the proposed method for different kinds of fusion between granularities for the PASCAL VOC and TRECVID datasets, respectively. Definition of fusions is as in Table 1. Based on Figures 9 and 10, with fusion of different granularities, performance usually increases. This shows the advantage of using multi-granularity partitioning and classification. When very small granularity (4×4 grid) is used for fusion, the performance decreases (especially for TRECVID dataset that has non-object concepts). This indicates that very small granularities cannot help for the detection even in the fusion.

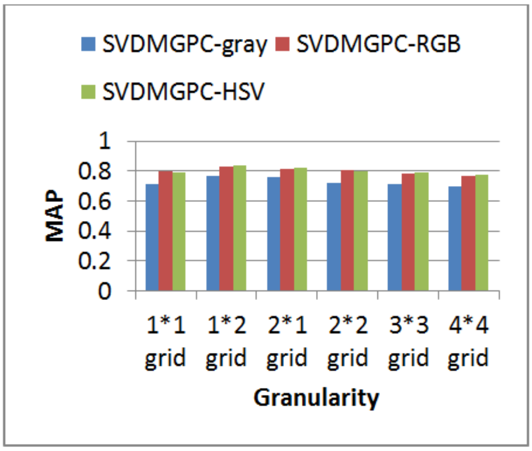


Fig. 7. MAP for different granularities for three cases of the proposed method for PASCAL VOC dataset.

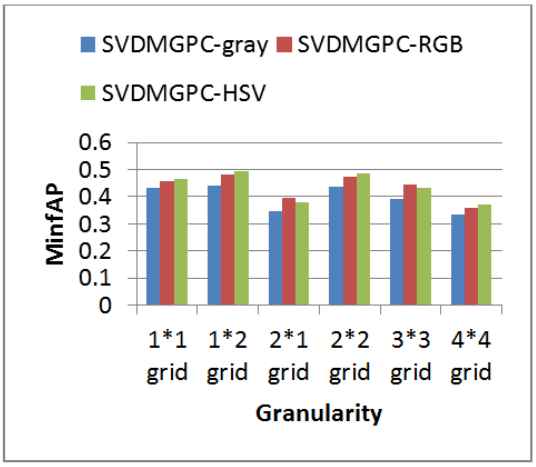


Fig. 8. MinfAP for different granularities for three cases of the proposed method for TRECVID dataset.

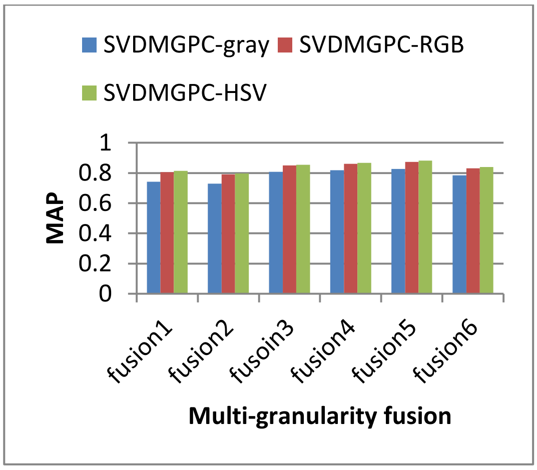


Fig. 9. MAP for different kinds of multi-granularity fusions for three cases of the proposed method for PASCAL VOC dataset.

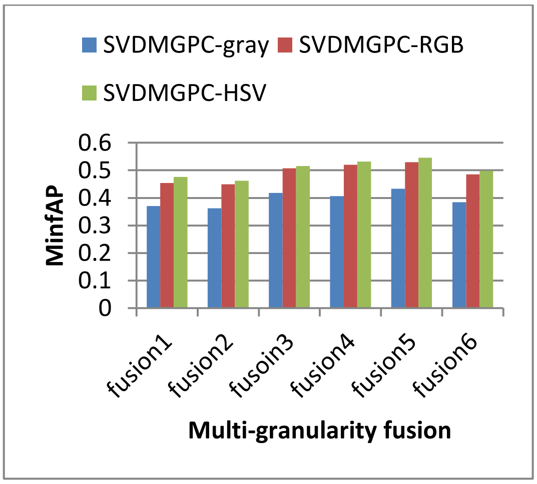


Fig. 10. MinfAP for different kinds of multi-granularity fusions for three cases of the proposed method for TRECVID dataset.

Now, we evaluate the proposed SVD features. For having a fair comparison in aspect of low-level features, widely-used local and global features in the literature, i.e. MPEG7 and BoW features, are selected for the comparison with the SVD features. Other referred works are not used for comparison. The reason is that either they are variations of SIFT features (like Fisher vector) with their limitations or they do not consider just static low-level features (e.g. CNN performs both low-level feature and classification together). Additionally, we want to use multiplicative distance in our method since dimensionality of SVD features is high. Therefore, for a fair comparison we must use methods that multiplicative distance is applicable to them, and for this reason CNN cannot not be used.

For the comparison, MPEG-7 visual features [3] including 81-d color moments, 64-d color histogram, 62-d homogeneous texture, 80-d edge direction histogram, and the local feature [4] with 128-d SIFT descriptor are used. The compared method with these features is represented with CCWES. For fair comparison these features are extracted with the proposed multi-granularity partitioning scheme on HSV color space. For each feature of each partition, classification is performed separately using the K-NN algorithm with the multiplicative distance. Selection of the control power for each kind of the feature vector, determining positive test images, and forming ranked lists of images are carried out similar to our method. The best case of fusion, i.e. fusion5, is selected for the comparison of the global and local features with the proposed SVD features. Therefore, the difference between CCWES and SVDMGPC is just in the low-level features.

The performances of the proposed and CCWES systems for PASCAL VOC and TRECVID datasets are reported in Figures 11 and 12, respectively. Table 2 also reports the overall performance of the proposed and CCWES methods. Based on Table 2, using the proposed SVD features (in SVDMGPC-HSV and SVDMGPC-RGB) gives the superior performance over using the common and global features (in CCWES) for both of datasets. Moreover, as it can be seen from Figures 11 and 12, SVDMGPC-HSV and SVDMGPC-RGB have better performance than CCWES for detecting most of concepts. The reason is that in SVD feature texture, color and edge information is stimulatingly integrated with considering their relationship, and this information is sorted in accordance with their importance for representing concepts. But, these properties cannot be captured in the compared features.

Furthermore, based on Table 2, SVD features in the color space gives better performance than SVD features in the gray scale space. This shows that the color information can have an important impact on the concept detection performance. However, for some concepts in the two datasets the SVD features in the gray scale space gives better results than the SVD features in the color space. This indicates that color information does not always help for detection since some concepts may not be dependent on specific colors. One weakness of the proposed SVD features is that images should be in the same size. For solving this problem all images should be resized to the same size. Of course, this problem does not exist for the global and local features.

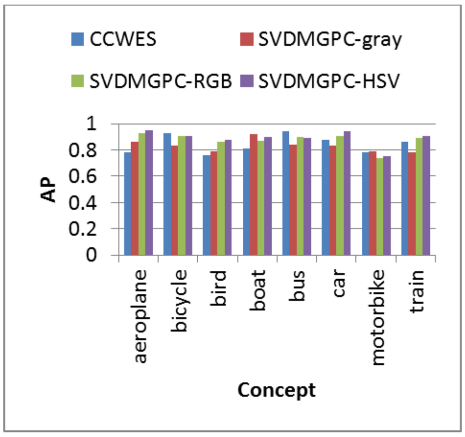


Fig. 11. AP for CCWES and three cases of the proposed methods for PASCAL VOC dataset.

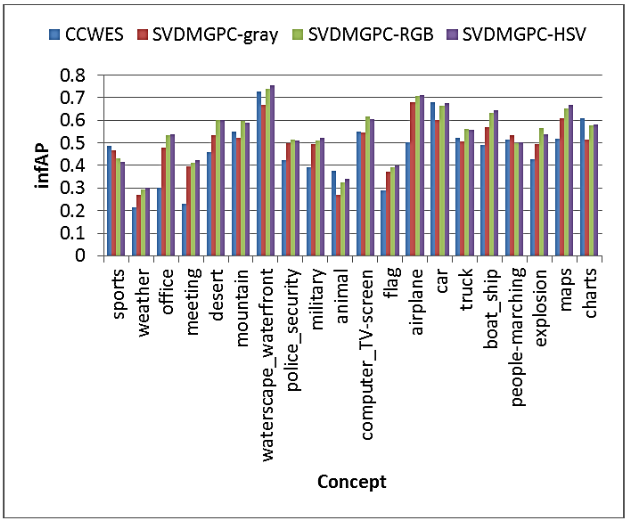


Fig. 12. infAP for CCWES and three cases of the proposed methods for TRECVID dataset.

Table 2. The overall performance of the CCWES and three cases of the proposed methods.

Methods	CCWES [3,4]	SVDMGPC-gray	SVDMGPC-RGB	SVDMGPC-HSV
MAP (PASCAL VOC)	0.8424	0.8263	0.8747	0.8820
MinfAP (TRECVID)	0.4637	0.4329	0.5297	0.5456

To confirm whether the improvement of the proposed method is statistically significant, we further conduct randomization test (suggested by TRECVID [32]) on the proposed and compared methods. In this test, the standard number of iterations in the randomization is 10000 and the standard level of significance is 0.05. The results of this test are shown in Table 3. P-value is the probability that the difference between two methods is due to chance. From these results, it is observed that for both of datasets the improvements of SVDMGPC-HSV and SVDMGPC-RGB over CCWES (and also SVDMGPC-gray) are statistically significant. Moreover, there is no significant difference between SVDMGPC-HSV and SVDMGPC-

RGB. Furthermore, the difference between SVDMGPC-gray and CCWES is not statistically significant.

Table 3. p-values of the significance test between methods for PASCAL VOC and TRECVID datasets.

Methods	p-value (PASCAL VOC)	p-value (TRECVID)
SVDMGPC-HSV vs. SVDMGPC-RGB	0.0706	0.1075
SVDMGPC-HSV vs. CCWES	0.0001	0.0332
SVDMGPC-HSV vs. SVDMGPC-gray	0.0005	0.0003
SVDMGPC-RGB vs. CCWES	0.0006	0.0397
SVDMGPC-RGB vs. SVDMGPC-gray	0.0014	0.0005
CCWES vs. SVDMGPC-gray	0.1976	0.4227

Total training and test time for the proposed and CCWES methods are shown in Table 4. The simulations have been carried out on a PC with Intel Core i7 CPU 2.79 GHz, and 8GB RAM with MATLAB. The proposed method consumes more training and test time. The reason is that extracting SVD features needs more time than extracting local and global features in CCWES. Moreover, SVD features have much more dimensionality than the features in CCWE and this leads to the more computations.

Table 4. Total training and test time (hours) for the proposed and compared methods for PASCAL VOC and TRECVID datasets.

Methods	CCWES [3,4]	SVDMGPC-gray	SVDMGPC-RGB	SVDMGPC-HSV
Training time (PASCAL VOC)	12.4	18.6	53.8	56.2
Test time (PASCAL VOC)	1.9	7.5	19.3	21.7
Training time (TRECVID)	51.6	75.0	212.7	216.4
Test time (TRECVID)	9.2	32.7	88.3	91.5

6. Conclusion

In this paper, new kind of static visual features, namely, right and left singular feature vectors and singular value feature vector, were proposed which were derived by applying SVD directly to the raw images. These features were different from features of methods in which SVD was applied to the low-level features of images (like in PCA or LAS techniques). Particularly, the proposed SVD features had this advantage that in which edge, color and texture information was integrated simultaneously and was sorted in accordance with their relationship and importance for the concept detection.

Additionally, feature extraction was performed in the multi-granularity manner. Furthermore, in the proposed method classification was carried out for each partition of each granularity separately, in contrast to the existing systems in which classification was performed for the whole image not for each partition. The proposed multi-granularity partitioning and classification had this advantage that the results of classifications on partitions with and without the target concept were not affected each other. This led to the performance improvement of the concept detection. Since usually feature vectors were high-dimensional even after the dimension reduction, classification was carried out by the K-NN algorithm with a new distance function, the multiplicative distance. This distance function was stable in the high-dimensional space, and was also usable in the low-dimensional space.

Experimental results showed the superiority of the multi-granularity partitioning and classification method over the spatial pyramid matching method and classification on the whole image and also the superiority of the proposed SVD features over the widely-used local and global features for the concept detection. However, the proposed method consumes more training and test time. The reason is that extracting SVD features needs more time than extracting local and global features in the compared method. Moreover, SVD features have much more dimensionality than the compared features, and this results in more computations.

References

[1] M. Jiu and H. Sahbi, “Nonlinear Deep Kernel Learning for Image Annotation,” IEEE Trans. Image Process., vol. 26, no. 4, pp. 1820-1832, Apr. 2017.

[2] H. Kaur and V. Dhir, “Local maximum edge cooccurance patterns for image indexing and retrieval,” Int’l Conf. Signal and information Processing, 2016.

[3] Moving Picture Expert Group. [Online]. Available: http://www.chiariglione.org/mpeg

[4] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. Hauptmann, “Representations of keypoint-based semantic concept detection: A comprehensive study,” IEEE Trans. Multimedia, vol. 12, no. 1, pp. 42–53, Jan. 2010.

[5] K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, “When is nearest neighbors meaningful?” in Proc. Seventh Int’l Conf. Databasse Theory (ICDT ’99), 1999, vol. 1540, pp. 217–235.

[6] C.-M. Hsu and M.-S. Chen, “On the design and applicability of distance functions in high-dimensional data space,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 4, pp. 523–536, Apr. 2009.

[7] R. J. Durrant and A. Kaban, “When is ‘nearest neighbour’ meaningful: A converse theorem and implications,” J. of Complexity, vol. 25, no. 4, pp. 385–397, 2009.

[8] J. Mansouri and M. Khademi, "Multiplicative Distance: A method to alleviate distance instability for high-dimensional data," Knowledge and Information Systems, vol. 45, no. 3, pp. 783-805, 2015.

[9] Y, Han, Y. Yang, Y. Yang, Z. Ma, N. Sebe, and X. Zhou, "Semisupervised feature selection via spline regression for video semantic recognition," IEEE Trans. Neural Networks and Learning Systems, vol. 26, no. 2, pp. 252–264, Feb. 2014,

[10] L. Duan, I. W. Tsang, and D. Xu, "Domain transfer multiple kernel learning," IEEE Trans. Pattern Anal. Machine Intell., vol. 34, no. 3, pp. 465–479, Mar. 2012.

[11] P. Srivastava and A. Khare, “Integration of wavelet transform, Local Binary Patterns and moments for content-based image retrieval,” Journal of Visual Communication and Image Representation, vol. 42, pp. 78-103, 2017.

[12] X. Zhang and C. Liu, “Image understanding based on histogram of contrast,” Signal, Image and Video Processing, vol. 10, no. 1, pp 103–112, 2016.

[13] S. Lazebnik, C. Schmid, and J. Ponce, “Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories,” Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, pp. 2169-2178, 2006.

[14] J. Sanchez and J. Redolfi, “Exponential family Fisher vector for image classification,” Pattern Recognition Letters, vol. 59, pp. 26-32, July 2015.

[15] K. Lim and H. Wang, “Sparse Coding Based Fisher Vector Using a Bayesian Approach,” IEEE Signal Processing Letters, vol. 24, no. 1, pp. 91-95, Jan. 2017.

[16] Y. Shi, Y. Wan, K. Wu and X. Chen, “Non-negativity and locality constrained Laplacian sparse coding for image classification,” Expert Systems with Applications, vol. 72, pp. 121-129, 2017.

[17] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, and S. Yan, “Cross-Modal Retrieval With CNN Visual Features: A New Baseline ,” IEEE Trans. Cybernetics, Vol. 47, no. 2, pp. 449-460, Feb. 2017.

[18] G. Wang and Q. M. J. Wu, Advances in Pattern Recognition: Guide to Three Dimensional Structure and Motion Factorization, London: Springer, 2011.

[19] H. Yanai, K. Takeuchi, and Y. Takane, Projection Matrices, Generalized Inverse Matrices, and Singular Value Decomposition, New York: Springer, 2011.

[20] M. Narwaria and W. Lin, "SVD-based quality metric for image and video using machine learning," IEEE Trans. Syst., Man, Cybern. B, Cybern., vol. 42, no. 2, pp. 347–364, Apr. 2012.

[21] M. Radovanovi'c, A. Nanopoulos, and M. Ivanovi´c, "On the existence of obstinate results in vector space models," in Proc. 33rd Int. ACM SIGIR conference on Research and development in information retrieval, New York, 2010, pp. 186–193.

[22] J. Hegde, "Time course of visual perception: Coarse-to-fine processing and beyond," Progress in Neurobiology, vol. 84, pp. 405–439, 2008.

[23] M.D. Menz and R.D. Freeman, "Stereoscopic depth processing in the visual cortex: A coarse-to-fine mechanism," Nat. Neurosci., vol. 6, pp. 59–65, 2003.

[24] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. (2007). The PASCAL Visual Object Classes Challenge Results [Online]. Available: http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html

[25] M. Sanderson, “Test collection based evaluation of information retrieval systems,” Foundations Trends Inform. Retrieval, vol. 4, no. 4, pp. 247–375, 2010.

[26] P. Over, G.M. Awad, W. Keraaij, and A.F. Smeaton, "TRECVID 2007-overview," in TRECVid 2007 - Text REtrieval Conference TRECVid Workshop, Gaithersburg, Maryland, Nov. 2007.

Kamran Farajzadeh is Ph.D student of IT management in Islamic Azad University, Science and Research Branch, Tehran, Iran. His major research interests are wireless sensor networks, Internet of Things, image processing, robotic, medical informatics and IT management.

Esmail Zarezadeh was born in Tehran, Iran. He received the B.Sc degree in electrical engineering, and M.Sc degrees in electrical engineering and space engineering in 2008, 2013, respectively. Now he is Ph.D Student in electrical engineering at the Amir Kabir University of Technology (Tehran Polytechnic), Tehran, Iran. His working experiences are Radar systems, digital communication, application of sensing in multi antenna network and Adhoc systems. His research interests are MTM, RF/microwave circuits design and image processing.

Jafar Mansouri received Ph.D of electrical/communication engineering at Ferdowsi University of Mashhad, Iran, in 2015. His main research interests include image and video processing and analysis, multimedia information retrieval, high-dimensional data analysis, machine learning and computer vision.