Pattern Recognition Letters 32 (2011) 1902–1911

Contents lists available at SciVerse ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Simple low-dimensional features approximating NCC-based image matching

Shin'ichi Satoh

National Institute of Informatics, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, Japan

ARTICLE INFO

Article history:
Received 8 April 2010
Available online 16 August 2011
Communicated by G. Borgefors

Keywords:
Image matching
Image mining
Normalized cross correlation

ABSTRACT

This paper proposes new low-dimensional image features that enable images to be very efficiently matched. Image matching is one of the key technologies for many vision-based applications, including template matching, block motion estimation, video compression, stereo vision, image/video near-duplicate detection, similarity join for image/video database, and so on. Normalized cross correlation (NCC) is one of widely used method for image matching with preferable characteristics such as robustness to intensity offsets and contrast changes, but it is computationally expensive. The proposed features, derived by the method of Lagrange multipliers, can provide upper-bounds of NCC as a simple dot product between two low-dimensional feature vectors. By using the proposed features, NCC-based image matching can be effectively accelerated. The matching performance with the proposed features is demonstrated using an image database obtained from actual broadcast videos. The new features are shown to outperform other methods: multilevel successive elimination algorithm (MSEA), discrete cosine transform (DCT) coefficients, and histograms, achieving very high precision while only slightly sacrificing recall.

1. Introduction

This paper proposes new low-dimensional image features that enable very efficient image matching. Image matching is one of the key technologies for many vision-based applications including template matching (Di Stefano and Mattoccia, 2003; Mattoccia et al., 2008), block motion estimation and video compression (Gao et al., 2000), and stereo vision (Hirschmüller and Scharstein, 2009). Recently new applications of image matching for large-scale image/video database are started to be explored, such as image near-duplicate detection (Douze et al., 2009), image mining and its application to image annotation (Wang et al., 2010). For example, in (Jing and Baluja, 2008), given a set of images returned by The Google Images search engine, pairs of similar images are detected and regarded as links between images, and then a modified version of PageRank is applied to extract images that are more relevant to the query. Video near-duplicate detection has recently been intensively studied, which can be regarded as matching similar videos within video archives. There is a research trend where near-duplicate shots (we can regard these as shots sharing the same video materials) are regarded as links within video archives, and the archives are then analyzed based on the structure of the links. Such attempts have included threading of news topic (Hsu and Chang, 2006), generation of auto documentaries (Wu et al., 2006), ranking of news topics (Zhai and Shah, 2005), mining and browsing of important shot (Yamagishi et al., 2004).

Well-known image matching methods are sum of absolute differences (SAD), sum of squared differences (SSD), and normalized cross correlation (NCC), which are known to be simple yet effective. Among these, NCC is one of widely used methods for image matching with preferable characteristics such as robustness to intensity offsets and contrast changes (Forsyth and Ponce, 2002; Ballard and Brown, 1982). However, NCC is sometimes computationally expensive especially when the size of images is large. In order to employ NCC as image matching to large-scale image/video database, intensive speedup of NCC calculation is needed.

To achieve this end, this paper introduces new types of features that enable images to be very efficiently matched. The proposed features are designed to provide upper-bounds of NCC as a simple dot product between two low-dimensional feature vectors. Since the features satisfy upper-bounding conditions, matching based on these can be used as a first-stage filter followed by evaluating NCC, without any false dismissals. In addition, we also found that the matching criteria using the features could be made slightly more stringent to precisely approximate NCC. As the upper-bound is obtained with the method of Lagrange multipliers, and its derivative is zero around the bound, the approximated NCC using the proposed features is very insensitive to small changes in image intensities. Therefore, raising the threshold for the approximated NCC could drastically improve precision with only a very small increase in the number of false dismissals. Compared to other well-known features such as multilevel successive elimination algorithm (MSEA) (Gao et al., 2000), DCT coefficients, and histograms, the proposed features could achieve very high precision while only slightly sacrificing recall. The matching performance

E-mail address: satoh@nii.ac.jp

0167-8655/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2011.07.027

of new features is demonstrated by an image database obtained from actual broadcast videos.

2. Related works

Image matching has been studied since the dawn of digital-image processing. A typical technique is template matching, which is aimed at finding and locating a small image region (patch) within a given image that matches the template image.

Key-point-based representations of local features have recently been studied in the field of image matching. The basic idea is first to detect a number of key points (a few hundreds to a few thousands) from an image, then represent the image with a set of local features each of which is extracted from the local vicinity of each key point. This representation is beneficial because it is robust against partial clipping or occlusion, and can handle sub-region-to-sub-region matching. The scale-invariant feature transform (SIFT) key point detector and local features (Lowe, 2004) have typically been used due to their invariance to scale and rotation, and their robustness against changes in lighting conditions and distortions by perspective projection. There have been a number of reports applying the key-point-based method to image matching (Lejsek et al., 2006; Law-To et al., 2006; Joly et al., 2007; Zhao et al., 2007). However, it is well-known that pairwise point matching is required to match images precisely, and this is obviously computationally very demanding (Zhao et al., 2007). On the other hand, holistic image matching using global feature is also employed for image/video database mining seeking for higher computational efficiency when robustness to occlusion and clipping is not seriously required. Berrani et al. (2008), Döhring and Lienhart (2009) and Wu et al. (2010) report commercial film mining methods from television video streams, and (Wang et al., 2010) applies image near-duplicate detection to 2 billion images in the Web for image auto-annotation.

Pixel-wise comparison has been used for a long time, especially for template matching. Typically, differences in intensities between corresponding pixels are computed and then aggregated into one value indicating the similarity between images. For example, the sum of the absolute distance (SAD) or the sum of the squared distance (SSD) are typically used for image matching (Hirschmüller and Scharstein, 2009). These methods have drawbacks in that they are less robust against geometrical distortion including that from occlusion, clipping, perspective projection, as well as illumination distortion. However, they are known to be very simple, are suitable for hardware implementation such as digital signal processor (DSP), and can match images sufficiently precisely for many applications such as stereo vision, block-based motion estimation and video compression, and template matching. Normalized cross correlation (NCC) has excellent properties due to its inherent normalization, viz., all images are equivalently normalized to zero mean images and are tolerant against offsets in intensity (Forsyth and Ponce, 2002; Ballard and Brown, 1982). They are also normalized to images having unit variance, and are thus tolerant against variation in amplitude. Due to its preferable properties, NCC is employed in many applications including template matching (Di Stefano and Mattoccia, 2003; Mattoccia et al., 2008), block motion estimation and video compression (Gao et al., 2000), and stereo vision (Hirschmüller and Scharstein, 2009). Couple of methods to accelerate NCC calculation have also been studied. Tsai and Lin (2003) used the following idea where NCC-based template matching can be regarded as a convolution operation, where the template is scanned over the entire image, and the matching result of each location can be regarded as the value of the function of the convolution at the location. Convolution operation in the spatial domain can be very efficiently accomplished by multiplication in the frequency domain (via FFT), and thus NCC calculation can be accelerated by converting them to the frequency domain. Although this technique is effective for scanning a template over an image, it is not effective for image-to-image matching because it requires additional FFT and inverse FFT operations. On the other hand (Gao et al., 2000) proposes multilevel successive elimination algorithm (MSEA) to accelerate NCC calculation. The method is based on the idea that the upper-bound of NCC can be derived from partial computation within subregions. Since partial statistics (sum of squared intensities within subregion) can be precomputed, MSEA can convert an image into low-dimensional vector (the number of dimensions is equal to the number of subregions), and thus the upper-bound of NCC can be efficiently computed. The upper-bound is derived by using Cauchy–Schwarz inequality, and there are a host of papers based on this idea (Di Stefano and Mattoccia, 2003; Mattoccia et al., 2008). In addition, MSEA is followed by its extension such as adaptive MSEA where adaptive subregion decomposition is used (Wei and Lai, 2007).

The dirty filtering technique (Agrawal et al., 1993) is known as an effective method of accelerating the retrieval of items in a data set that match a certain criteria, especially when retrieval incurs costly distance calculations. The basic idea is to convert the original distance calculation into a computationally light distance calculation. The conversion should have a lower-bounding condition, where the converted distance is always lower than the original distance, so that the data to be retrieved are not missed. However, it is sometimes rather difficult to obtain a lower-bounding condition for conversion with sufficient efficiency. Linear projection such as the discrete cosine transform (DCT) and principal component analysis (PCA), and image matching by using Euclidean distance in the converted space is known to yield the lower-bounding condition. However, PCA requires training data and the results may depend on their distribution. DCT, on the other hand, is known to be a good approximation of PCA trained with a typical image distribution (Clarke, 1981). Histograms are sometimes used for rapid filtering (Agrawal et al., 1993; Yamagishi et al., 2003; Satoh, 2004). These can achieve good accuracy with features of very low dimension, but they do not normally satisfy lower-bounding conditions. This paper introduces a new way of representing features, which can be used as dirty filtering for NCC, while rigorously satisfying the lower-bounding condition (in our case the new features provide the upper bound of NCC, because NCC is similarity function, not distance function).

3. Problem description: matching image pairs within image database

Assume that two sets of images are given, $IS^1 = \left\{I_i^1\right\}$, and $IS^2 = \left\{I_j^2\right\}$. Assume further that NCC is used to check the similarity of images. The problem here is to detect all pairs between two image sets whose NCC values are above a certain threshold, $-1 \ll \theta < 1$. The result is:

$$\mathrm{IIPS} = \{(I^1, I^2) | I^1 \in IS^1,\ I^2 \in IS^2,\ \mathrm{NCC}(I^1, I^2) > \theta\}, \tag{1}$$

where NCC is defined as:

$$\mathrm{NCC}(I^1, I^2) = \frac{\sum_{x,y}(I^1 - \overline{I^1})(I^2 - \overline{I^2})}{\sqrt{\sum_{x,y}(I^1 - \overline{I^1})^2 \sum_{x,y}(I^2 - \overline{I^2})}}, \tag{2}$$

where $\bar{I}$ is the mean of I. A simple implementation is to check NCC for all combination of image pairs between image sets, viz., linear search. However, since NCC calculation is computationally demanding, simple linear search described above is not a practical solution. This is especially significant in matching images in large-scale image and video archives.

Dirty filtering techniques are known to be effective in such cases (Agrawal et al., 1993); addressing efficient searches of databases that require costly similarity (or distance) calculations. In the dirty filtering scheme, we first convert each element in the database (normally high-dimension: 10,000 or more) into a low-dimensional vector (tens to hundreds of dimensions) by a certain function *f*. By this conversion, the calculation of the metrics (distance, similarity, etc.) become much easier. For instance, the calculation of Euclidean distance is basically proportional to the number of dimensions. In addition, we can use multi-dimensional indexing techniques (Böhm et al. (2001) is a good survey) to efficiently search for low-dimensional data. *IIPS* is approximated as

$$\widetilde{IIPS} = \left\{(I^1, I^2) | I^1 \in IS^1,\ I^2 \in IS^2,\ s(f(I^1),\ f(I^2)) > \theta\right\}, \tag{3}$$

where $s(\cdot)$ represents the similarity between vectors and $\widetilde{IIPS}$ is the approximation of *IIPS*. Finally, each element of $\widetilde{IIPS}$ is "cleansed" by the original metric (NCC in our case), viz., pairs of images, whose similarities in the low-dimensional space are higher than the threshold but similarities in the original space are lower than the threshold, are removed. "The lower-bound condition," viz., $\mathrm{NCC}(I^1, I^2) > \theta \rightarrow s(f(I^1),\ f(I^2)) > \theta$, ensures that dirty filtering will not cause false dismissals, viz., $IIPS \subseteq \widetilde{IIPS}$. If the upper-bound of NCC is below θ, the corresponding image pairs can safely be discarded without causing false dismissals, and $s(f(I^1), f(I^2))$ gives the upper-bound of NCC. Note that NCC should be upper-bounded because NCC defines similarity but not distance. However, due to the discrepancy between the original metric and the metric in the converted low-dimensional space, this condition tends to be difficult to satisfy. If we somehow achieve $IIPS \subseteq \widetilde{IIPS}$, $\widetilde{IIPS}$ may become unexpectedly large, thus cleansing may take an impractically long time.

The effectiveness of the low-dimensional conversion and similarity can be evaluated using precision and recall:

$$\text{precision} = \frac{|IIPS \cap \widetilde{IIPS}|}{|\widetilde{IIPS}|}, \tag{4}$$

$$\text{recall} = \frac{|IIPS \cap \widetilde{IIPS}|}{|IIPS|}. \tag{5}$$

If the lower-bounding condition is satisfied, recall is equal to one. However, we can easily achieve recall = 1 by making $\widetilde{IIPS}$ very large (extreme case is $\widetilde{IIPS} = $ universal set). However this makes precision very low. Preferable low-dimensional conversion and similarity should achieve recall = 1 while making precision as high as possible. If recall is allowed to be less than one, both precision and recall should be made as high as possible. Thus, the issue is how to obtain the optimal function, *f*, which can achieve very high precision and recall.

4. Proposed features for image matching

We introduce new image features for image matching. The features are first introduced to obtain the upper-bound of NCC, as a simple dot product between two low-dimensional feature vectors. The features are based on the block decomposition of images, and thus can take full advantage of the Markovian properties of images, where nearby pixels tend to have similar intensities. After this, even though the matching criteria are made slightly more stringent, viz., the threshold in the low-dimensional space is made tighter than required in the original space, it is shown that the features still provide very precise approximation of NCC.

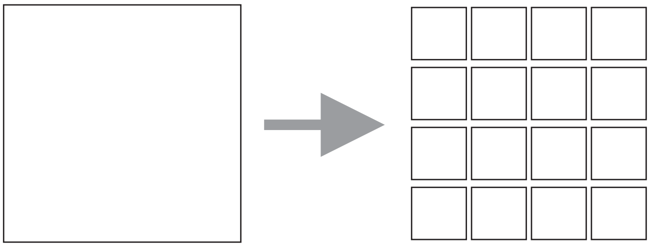

Fig. 1. Typical image decomposition by using rectangles of fixed size.

4.1. Upper- and lower-bounds of normalized cross correlation

Let us assume that an image is composed of a fixed number of pixels (discrete image representation). Let us further assume that the pixels comprising the image are decomposed into pixel groups each of which is composed of a fixed number of pixels. For instance, image x is composed of pixels $x_{i,j}$, $i = 1, \ldots, n$ and $j = 1, \ldots, m$, where the image is composed of n groups x_i, $i = 1, \ldots, n$ each of which is composed of m pixels $x_i = \{x_{i,j} \mid j = 1, \ldots, m\}$. Typically, such decomposition can be accomplished by dividing an image into rectangular regions of fixed size as shown in Fig. 1.

We first obtain NCC between two images, $\tilde{x}_{i,j}$ and $\tilde{y}_{i,j}$, $i = 1, \cdots, n$ and $j = 1, \cdots, m$, which represent pixel intensities. Normalized intensities $x_{i,j}$ can then be obtained by[1]:

$$x_{i,j} = \frac{\tilde{x}_{i,j} - \bar{\tilde{x}}}{\sqrt{\sum_{i,j}\left(\tilde{x}_{i,j} - \bar{\tilde{x}}\right)^2}}. \tag{6}$$

Then, the NCC between two images can be obtained by the sum of the products of the corresponding components of normalized intensities:

$$\mathrm{NCC}(x, y) = \sum_{i,j} x_{i,j} y_{i,j}. \tag{7}$$

Having obtained the notation, we can then study the upper- and lower-bounds of NCC between the images knowing the means (or sums) and variances (or squared sums) of the normalized intensities of pixels in blocks. That is,

$$X_i = \sum_j x_{i,j}, \tag{8}$$

$$XX_i = \sum_j x_{i,j}^2, \tag{9}$$

$$Y_i = \sum_j y_{i,j}, \tag{10}$$

$$YY_i = \sum_j y_{i,j}^2, \tag{11}$$

are assumed to be known where X_i (Y_i) is the sum of the normalized intensities of the ith group of x (y), and XX_i (YY_i) is the squared sum of the normalized intensities of the ith group of x (y). Due to the nature of normalized intensity, the following equations always hold:

$$\sum_i X_i = 0, \tag{12}$$

$$\sum_i XX_i = 1 \tag{13}$$

(the similar equations hold for y as well). Therefore, to obtain the upper- and lower-bounds of NCC, we need to minimize (or

[1] Strictly speaking, normalized intensities should be $x_{i,j} = \frac{\tilde{x}_{i,j} - \bar{\tilde{x}}}{\sqrt{\frac{1}{mn}\sum_{i,j}\left(\tilde{x}_{i,j} - \bar{\tilde{x}}\right)^2}}$, however, for notational simplicity, we use the definition here.

maximize) NCC(*x*,*y*) subject to $X_i = \sum_j x_{i,j}$, $XX_i = \sum_j x_{i,j}^2$, $Y_i = \sum_j y_{i,j}$, and $YY_i = \sum_j y_{i,j}^2$ for all *i*.

This can be solved by using the method of Lagrange multipliers. Let

$$\begin{aligned}\Lambda = \mathrm{NCC}(x,y) &- \sum_i \lambda_{1,i}\left(\sum_j x_{i,j} - X_i\right) \\ &- \sum_i \lambda_{2,i}\left(\sum_j x_{i,j}^2 - XX_i\right) - \sum_i \lambda_{3,i}\left(\sum_j y_{i,j} - Y_i\right) \\ &- \sum_i \lambda_{4,i}\left(\sum_j y_{i,j}^2 - YY_i\right), \end{aligned} \tag{14}$$

where $\lambda_{1,i}$, $\lambda_{2,i}$, $\lambda_{3,i}$, and $\lambda_{4,i}$ are 4*n* Lagrange multipliers. By setting the derivative $d\Lambda = 0$, we can obtain 2 mn equations ($\partial\Lambda/\partial x_{i,j} = 0$ and $\partial\Lambda/\partial y_{i,j} = 0$ for all *i* and *j*), and by solving the equations, we can obtain the upper- and lower-bounds of NCC as:

$$UB = \sum_i \left(\frac{X_i Y_i}{m} + \sqrt{\left(XX_i - \frac{X_i^2}{m}\right)\left(YY_i - \frac{Y_i^2}{m}\right)}\right), \tag{15}$$

$$LB = \sum_i \left(\frac{X_i Y_i}{m} - \sqrt{\left(XX_i - \frac{X_i^2}{m}\right)\left(YY_i - \frac{Y_i^2}{m}\right)}\right) \tag{16}$$

(the complete derivation is given in Appendix A). The upper-bound can further be converted as:

$$UB = \xi_x \cdot \xi_y, \tag{17}$$

where

$$\xi_x = \left[\frac{X_1}{\sqrt{m}} \cdots \frac{X_n}{\sqrt{m}} \sqrt{XX_1 - \frac{X_1^2}{m}} \cdots \sqrt{XX_n - \frac{X_n^2}{m}}\right]^T. \tag{18}$$

Thus, by converting image *x* into 2*n* element vector ξ_x, the upper-bound of NCC between two images *x* and *y* can be calculated by the dot product between the two corresponding vectors ξ_x and ξ_y. The lower-bound can also be represented similarly but the special attention must be paid to the difference in the sign (see Eqs. (15) and (16)). Now it is easy to show that the norm of ξ_x (and ξ_y) is equal to one:

$$|\xi_x|^2 = \sum_i \frac{X_i^2}{m} + \sum_i \left(XX_i - \frac{X_i^2}{m}\right) = \sum_i XX_i = 1, \tag{19}$$

by using Eq. (13). The dot product between vectors ξ_x and ξ_y can be converted into the Euclidean distance between the vectors:

$$|\xi_x - \xi_y|^2 = 2 - 2\xi_x^T \xi_y. \tag{20}$$

Thus, searching images *y* having a larger upper-bound for NCC with given image *x* can be accomplished by a range search in the Euclidean space of the converted vector, ξ_y and ξ_x. For instance, searching images *y* under the condition NCC(*x*,*y*) > θ can be done with a range search from ξ_x under condition $|\xi_x - \xi_y|^2 < 2 - 2\theta$. This is compatible with many existing indexing structures for high-dimensional data because they support range query with the Euclidean distance. Therefore, image matching using the proposed features can further be accelerated by the combination with indexing structures.

4.2. Tightness of bounds

Eq. (18) represents features to calculate the upper-bound of NCC, especially, if the number of groups in an image is small, the features are low-dimensional vectors. However, the properties of the features depend on the configuration of groups of pixels. To make the features accurately approximate NCC, the bound needs to be tight. The tightness of the bound can be evaluated by the difference between the upper- and lower-bounds:

$$UB - LB = 2\sum_i \sqrt{\left(XX_i - \frac{X_i^2}{m}\right)\left(YY_i - \frac{Y_i^2}{m}\right)} \tag{21}$$

and the first element within the square root is:

$$\begin{aligned} XX_i - \frac{X_i^2}{m} &= \sum_j x_{i,j}^2 - \frac{\left(\sum_j x_{i,j}\right)^2}{m} = \sum_j x_{i,j}\left(x_{i,j} - \frac{\sum_k x_{i,k}}{m}\right) \\ &= \frac{1}{m}\sum_j \sum_k \left(x_{i,j}x_{i,j} - x_{i,j}x_{i,k}\right) = \frac{1}{m}\sum_j \sum_{k<j} \left(x_{i,j} - x_{i,k}\right)^2 \\ &= \sum_j \left(x_{i,j} - \mathrm{E}_k(x_{i,k})\right)^2 = m\mathrm{Var}_i(x_{i,j}), \end{aligned} \tag{22}$$

where E_k gives an expectation over *k*, and $\mathrm{Var}_i(x_{i,j})$ gives the variance of $x_{i,j}$ within x_i. Similarly, the second element within the square root of Eq. (21) can be converted, and thus

$$UB - LB = 2m\sum_i \sqrt{\mathrm{Var}_i(x_{i,j})\mathrm{Var}_i(y_{i,j})}. \tag{23}$$

This can obviously be minimized in two ways. The first is by minimizing the variance of the normalized intensities within each group, and the second is by minimizing *m*. Since *m* is the number of pixels in one group, this can be minimized by using a small number of pixels in a group; however, this obviously incurs higher number of groups *n*, and thus high dimensional features, ξ_x and ξ_y, which is not practical. Minimizing the variance of the normalized intensities within each group, on the other hand, can be achieved by adopting appropriate configurations for groups of pixels, such that the difference in the normalized intensities between pixels within the same group is as small as possible. A well-known property of images is that the (normalized) intensity values of nearby pixels tend to be very close due to their Markovian nature. To make pixels within each group as close together as possible, an image can be decomposed into rectangular blocks, where preferably each can be close to a square, and these blocks can be regarded as groups. We employ this configuration. By changing the number of groups in an image, we could change the dimensions of the features.

4.3. Comparison with the state of the art: MSEA

We then compare the proposed method with the state of the art acceleration method: multilevel successive elimination algorithm (MSEA) (Gao et al., 2000). Both methods provide the upper-bound of NCC based on partial computation of regions (subregions or blocks). The key difference is their mathematical derivation; the proposed method derives the upper-bound based on the method of Lagrange multipliers, whereas MSEA derives the upper-bound based on Cauchy–Schwarz inequality. Applying Cauchy–Schwarz inequality given below:

$$\sum_i a_i \cdot b_i \leqslant \sqrt{\sum_i a_i^2} \cdot \sqrt{\sum_i b_i^2} \tag{24}$$

and following the block decomposition used for the proposed method, the upper-bound based on MSEA can be given as follows:

$$UB^{\mathrm{MSEA}} = \sum_i \sqrt{XX_i}\sqrt{YY_i}. \tag{25}$$

Obviously, by using the following feature vector representation,

$$\xi_x^{\mathrm{MSEA}} = \left[\sqrt{XX_1}\sqrt{XX_2} \cdots \sqrt{XX_n}\right]^T, \tag{26}$$

we can obtain

$$UB^{\mathrm{MSEA}} = \xi_x^{\mathrm{MSEA}} \cdot \xi_y^{\mathrm{MSEA}}. \tag{27}$$

Comparing this with the upper-bound of the proposed method (Eq. (15)), it is easy to show that

$$UB \leqslant UB^{\mathrm{MSEA}} \tag{28}$$

i.e., the proposed method gives strictly tighter upper-bound than MSEA (the derivation is given in Appendix B). Note that this holds when the proposed method and MSEA use the same block decomposition for images. In this case, the number of dimensions of features of the proposed method is twice as large as MSEA. In the next section, it will be shown that the proposed method gives tighter upper-bound than MSEA even when the same number of dimensions is used.

5. Experiments

5.1. Evaluation of the Accuracy as Approximated NCC

In this section, we discuss our study on the behavior of filtering in image matching using the proposed features. Our goal here is to enumerate all image pairs whose NCC is greater than θ, viz., $IIPS = \{(I^1, I^2) | I^1 \in IS^1, I^2 \in IS^2,\ \mathrm{NCC}(I^1, I^2) > \theta\}$. If we use the dot product (similarity) based on the proposed features as the upper-bound and the same threshold value θ (given for NCC) as a filter, we will not miss any image pairs with NCC greater than $\theta : II\tilde{P}S = \{(I^1, I^2) | I^1 \in IS^1, I^2 \in IS^2, UB(I^1, I^2) > \theta\}$. The lower-bounding condition is satisfied by using this, and we can achieve 100% recall. (Note that the original lower-bounding condition is defined for distance, i.e., two images are matched if the distance is small, while NCC is large. Therefore, the upper-bound of NCC is used for the lower-bounding condition.)

To observe the behavior of features with actual images, we prepared two sets of 10,000 frame images (352 × 240 pixels) randomly selected from two different 1-h videos (total of 20,000 images). These 1-h videos were broadcast in the same time slot and on the same channel, but on different days. Since they were taken from the same time slot, they included the same video programs (actually the slot included news and a documentary), and thus each image set was expected to include some identical shots, such as opening and ending shots. Examples of identical shots are shown in Fig. 2. We plotted the precision by changing the dimensions of the features (Fig. 3), where θ was set to 0.9. For comparison, we conducted similar experiments by using DCT coefficients as features and the Euclidean distance as the distance satisfying the lower-bounding condition. We first converted the whole normalized image into frequency domain by using DCT, and resultant DCT coefficients were reordered by the zig-zag scanning used in JPEG and MPEG compression. The DCT-based low-dimensional features were then composed by selecting low frequency components of DCT coefficients. DC components were always zero and thus omitted, because of normalized images. We also compared with MSEA explained in Section 4.3. As Fig. 3 shows, the precision we achieved was not necessarily high, especially when the dimensions were low. The best precision was only 15% or less when the features had 1024 dimensions, while the precision fell to 5–6% with 64-D. Moreover, the DCT features achieved almost the same precision compared to the proposed features. MSEA resulted in significantly lower precision.

Even though the features we used could achieve 100% recall, the resulting precision was not necessarily high, and may have incurred heavier filtering at a later stage. However, in some cases, we could have raised the threshold for the upper-bound for NCC (giving up 100% recall). If sufficiently high precision and recall can be achieved, the output of the dirty filter can be used as an approximate result without precise filtering. This could be achieved by using a higher threshold for dirty filtering, i.e., $\theta \leqslant \theta'$. The precision and recall curves were plotted for features of different lengths by changing θ' (Fig. 4). In addition, three types of features were also compared, i.e., DCT, MSEA, and normalized intensity histograms (NIH) (Yamagishi et al., 2003; Satoh, 2004). NIHs with Euclidean distance do not satisfy the lower-bounding condition, and thus cannot ensure 100% recall, but are known to achieve very high precision and recall. An NIH was obtained by first decomposing each image into sub-regions (2 × 2, 3 × 3, etc.), and a histogram of normalized intensity was calculated for each sub-region by quantizing the domain of normalized intensity. A feature vector was composed by concatinating all histograms of

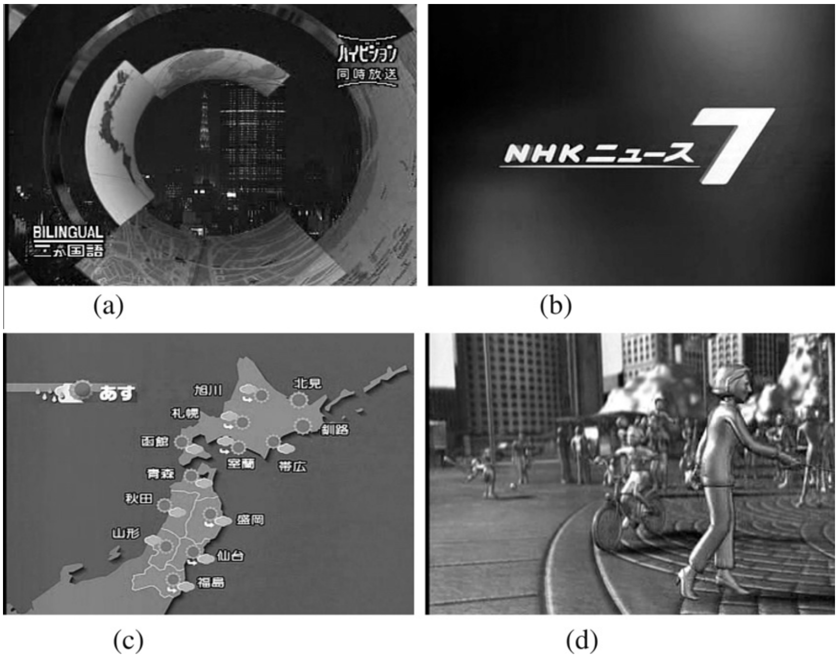

Fig. 2. Examples of identical shots. (a) Program opening. (b) Program title. (c) Weather chart. (d) CG shot.

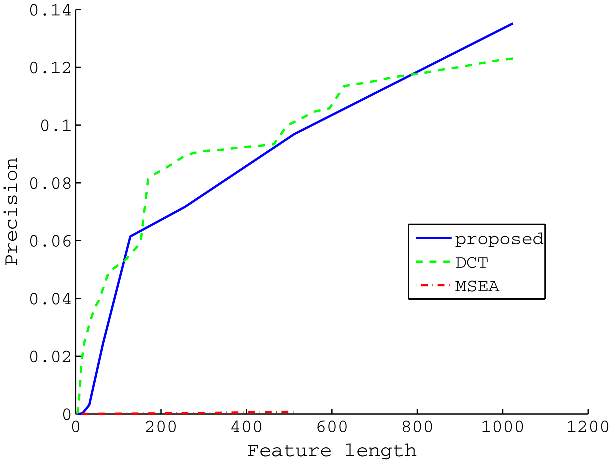

Fig. 3. Precision assuring perfect recall.

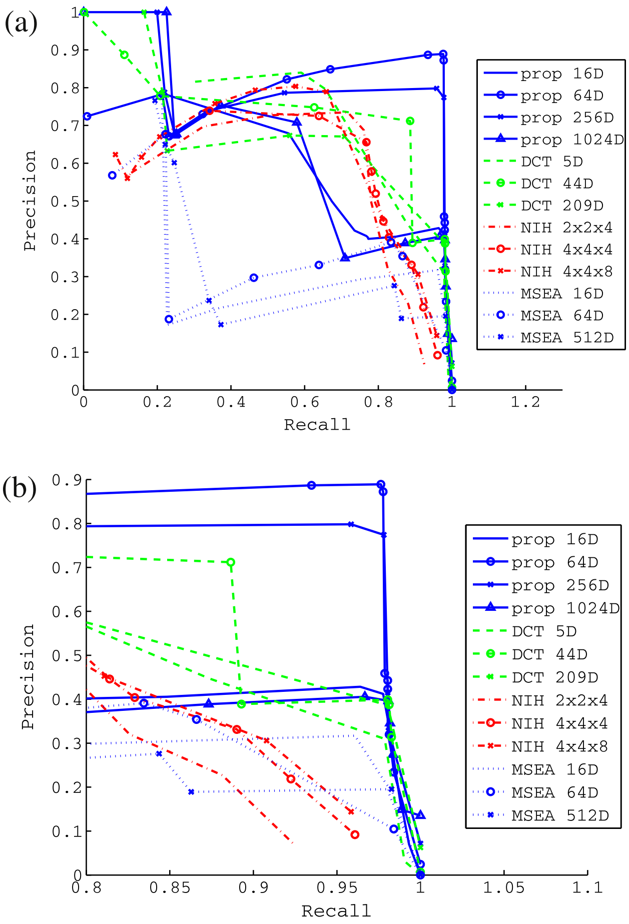

Fig. 4. Precision and recall curves. (a) Full. (b) Zoomed (recall rate above 0.8).

sug-regions. A different number of dimensions for features was achieved by changing the quantization and number of subregions. For example, NIH using decomposition of 2 × 2 subregions and 4-level quantization to normalized intensity was denoted as NIH 2 × 2 × 4.

Fig. 4 shows that the proposed features achieved more than 70% precision while retaining 97% recall with 64–256 dimensions. Especially when recall was very high (more than 95%), the proposed features achieved higher precision than DCT, NIH, or MSEA.

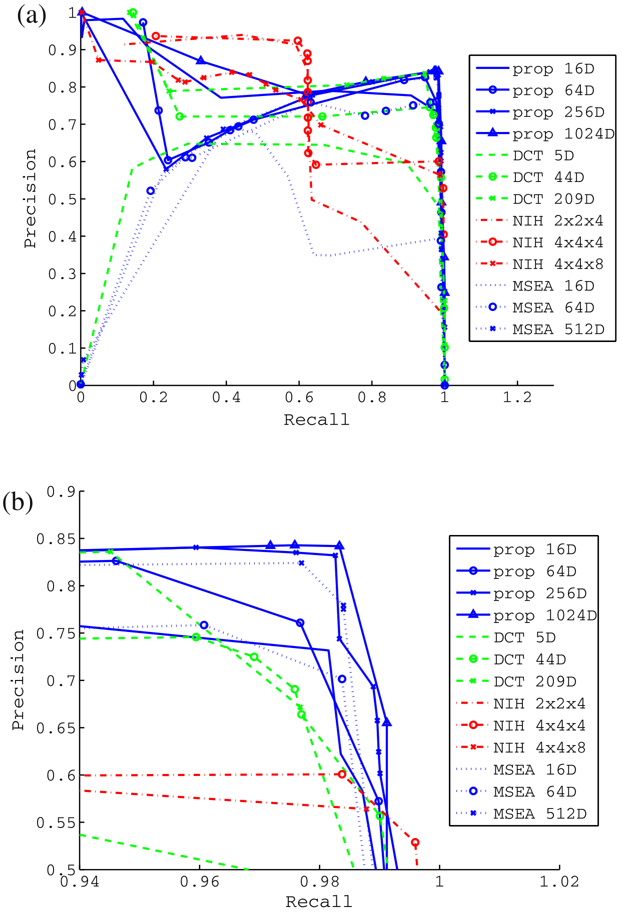

Fig. 5. Precision and recall curves with another pair of image sets. (a) Full. (b) Zoomed (recall above 0.94, precision above 0.5).

Fig. 5 shows the similar graph but using a different pair of image sets (10,000 images each) similarly obtained from broadcasted videos. Compared to Fig. 4, relatively higher precision and recall were achieved by almost all types of features. If we further look at the zoomed graph (Fig. 5b), we can observe that the proposed features of 1024D and 256D achieved higher performance, MSEA 512D achieved the next higher performance, the proposed features of 64D and 16D achieved the next, MSEA 64D followed, and the performance of DCT features was inferior to the above features. In summary, all configurations of the proposed features achieved stably higher performance than the others. For instance, the performance of MSEA 16D was significantly worse than the proposed features (including the proposed method with 16D configuration).

To further illustrate the advantage of the proposed features, precision has been plotted at fixed recall. Fig. 6 shows the precision of the proposed features compared with DCT, NIH, and MSEA. The proposed features clearly outperformed the other features with very high recall. When recall = 0.99, the proposed features achieved slightly worse performance than DCT, however, the precision is rather very low, namely, up to 0.14. We can observe that, in some cases, the proposed method performed better using fewer dimensions (see, for instance, Fig. 6a for recall = 0.95: the precision of 64D (128D, 256D) was better than that of 512D). Similar phenomenon is observed with MSEA (Fig. 6a: the precision of 16D (32D, 64D) is better than of 128D). Since both methods are based on block-based subregion decomposition, this phenomenon can be explained by the existence of an optimal size of blocks to achieve good precision when non-perfect recall setting is used. From Fig. 6, it could be observed that the proposed method achieved the best precision when 64 dimensional features were

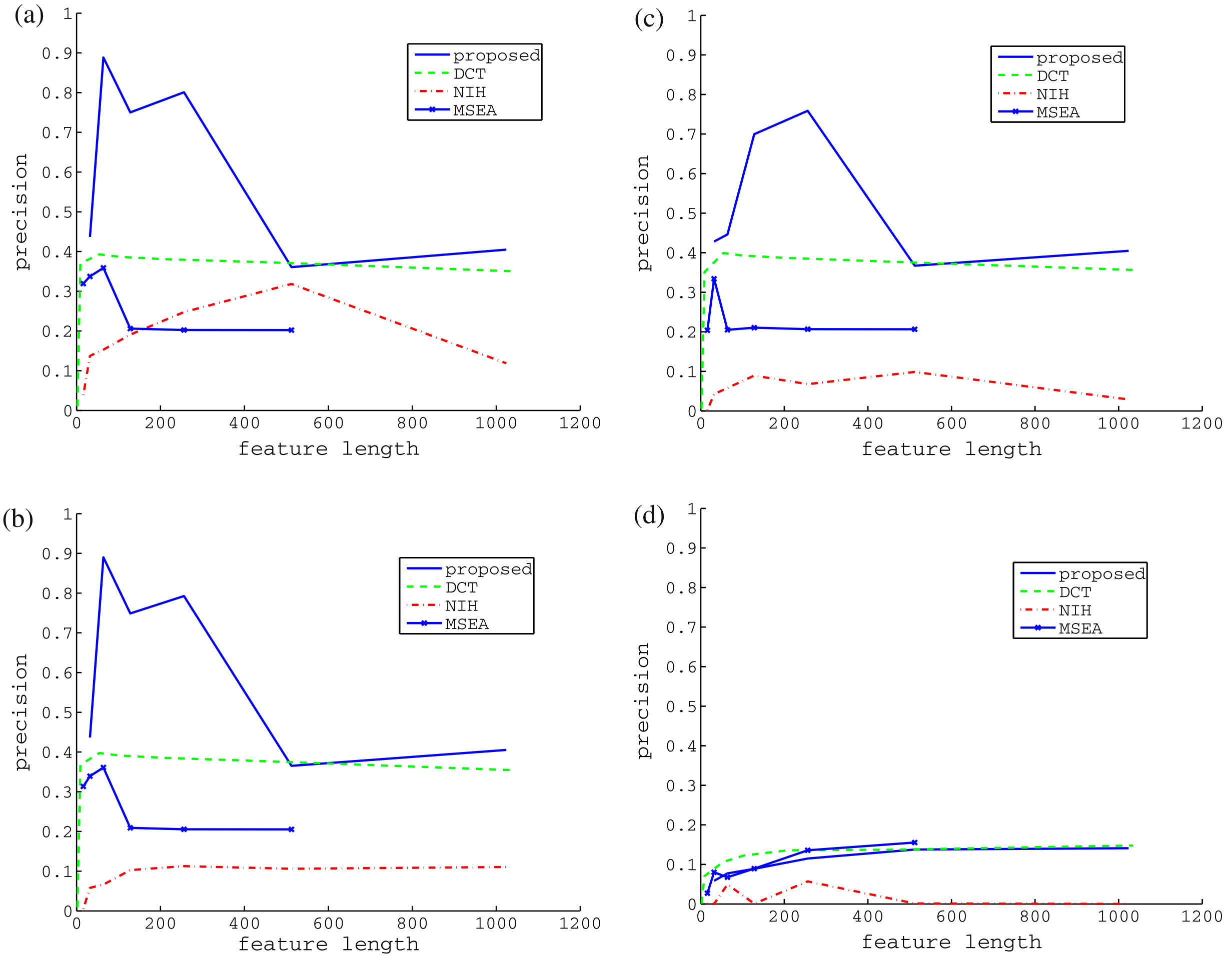

Fig. 6. Precision at fixed recall (note that the curve is Fig. 3 when recall = 1). (a) Recall = 0.95. (b) Recall = 0.97. (c) Recall = 0.98. (d) Recall = 0.99.

used, where the decomposition of 32 subregions was used. While MSEA achieved the best precision when 64 dimensional features, and thus the decomposition of 64 subregions, were used.

Fig. 7 shows the similar experiments using the other pair of image sets. Note that the precision was relatively high for almost all feature lengths. Even though recall was fixed to 0.99, the precision was around 0.6 (see Fig. 6d for comparison). Fig. 7a–c shows that the proposed feature and MSEA achieved better performance than DCT and NIH when relatively low recall (0.95–0.98) was used, while Fig. 7d shows that the proposed feature, DCT, and NIH achieved better performance than MSEA. The experiments showed that the proposed feature achieved consistently better performance than other features.

The proposed features were derived from the upper-bound of NCC by using the method of Lagrange multipliers. This makes the derivative of (approximated) NCC very close to zero when NCC as well as the approximated NCC are close to the threshold (θ). However, if NCC is not close to the threshold, the derivative is not close to zero. Inherently, the features can achieve 100% recall (when both NCC and approximate NCC are above the threshold). Even when the threshold is slightly increased, recall will not be greatly affected because the derivative is close to zero (when NCC is close to the threshold due to high recall), while precision is drastically increased because the derivative is not necessarily close to zero (when NCC is not necessarily close to the threshold). This could be one of the reasons the proposed features can achieve very high precision while retaining very high recall.

5.2. Evaluation of computational efficiency

We then evaluated the computational efficiency of image matching using the proposed features. Note that we used simple exhaustive search, i.e., distances of all possible pairs of vectors between two image sets were computed, to better demonstrate the effectiveness of the approximation of NCC by the proposed method. The search is essentially k-NN search in high-dimensional space, and thus can easily be accelerated by many existing methods including tree-based method such as ANN (Arya et al., 1998), hash-based method such as LSH (Andoni and Indyk, 2008), hierarchical vector quantization (Nister and Stewenius, 2006), vector quantization with scalar quantization (hamming embedding) (Douze et al., 2009), and so on, but at the cost of approximated results.

We tested two types of filter configurations for image matching. The first was a dirty filter only, where approximated NCC with the proposed features was used to filter image pairs, but without NCC-based cleansing. The second was a dirty filter and NCC, where dirty filtering was followed by precise comparison of image pairs by using NCC. The threshold for dirty filtering was also changed. First, the same threshold as that for NCC was used for dirty filtering (in our case $\theta = 0.9$), and thus 100% recall was ensured. Secondly, higher threshold than the NCC threshold was used so that to make recall 95%. The same sets of images were used. The computation time as well as the precision and recall were measured for several types of filter configurations.

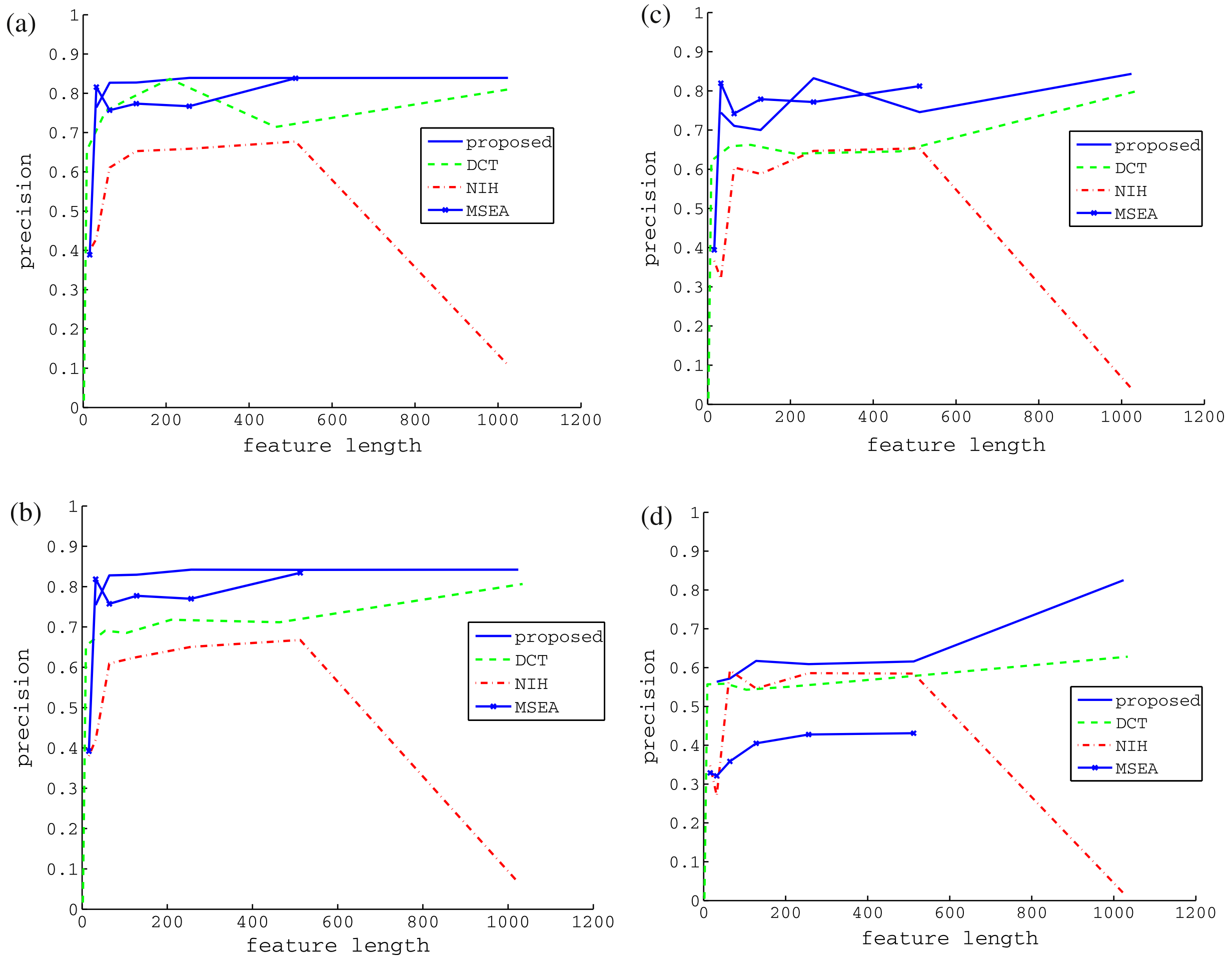

Fig. 7. Precision at fixed recall with another pair of image sets. (a) Recall = 0.95. (b) Recall = 0.97. (c) Recall = 0.98. (d) Recall = 0.99.

Table 1
Computation time.

Filter-type	Prec.	Recl.	Time (s)
32	0.0308	1	97.1
64	0.0243	1	144
128	0.0615	1	233
1024	0.135	1	1560
NCC	1	1	≈40 [h]
32-NCC	1	1	5580
64-NCC	1	1	846
128-NCC	1	1	507
1024-NCC	1	1	1690
32*	0.436	0.950	97.1
64*	0.888	0.950	144
128*	0.750	0.950	229
1024*	0.405	0.950	1560
32*-NCC	1	0.950	134
64*-NCC	1	0.950	163
128*-NCC	1	0.950	256
1024*-NCC	1	0.950	1600

The results are listed in Table 1. The numerals in the filter-type column mean the number of dimensions the features had, and the numerals followed by NCC mean dirty filter plus NCC filtering. The numerals with asterisks mean a dirty filter with a modified threshold at recall = 0.95. The time field is the wall clock time, except for NCC (estimated time is shown).

We can observe that the time spent by dirty filtering is proportional to the number of dimensions of the features. Based on this fact, the time required for NCC calculation was estimated (the number of dimensions is 352 × 240). The configuration of a dirty filter only with 100% recall resulted in rather low precision, as expected. However, if we combined a dirty filter with NCC, perfect results (100% precision and 100% recall) were accomplished. However, due to the low precision, many false alarms should be removed by NCC filtering, and the computation time was greatly increased. When the threshold for dirty filtering was changed, although the recall was reduced to 95%, the precision was drastically increased. Due to this, combination with NCC filtering did not greatly increase the computation time. Therefore, the configuration with slightly reduced recall (95%) with or without NCC filtering is thought to be a reasonable choice.

The speed accomplished by the proposed method may not be very high. However, please note that the proposed method is compatible with many acceleration techniques for multi-dimensional k-NN search (Arya et al., 1998; Andoni and Indyk, 2008; Nister and Stewenius, 2006; Douze et al., 2009). Namely, any multi-dimensional indexing methods including ANN (Arya et al., 1998), LSH (Andoni and Indyk, 2008), hierarchical vector quantization (Nister and Stewenius, 2006), etc., can be applied to a set of multi-dimensional vectors obtained by our proposed method, in place of simple exhaustive search used in the experiments, to realize accelerated k-NN search. Moreover, the proposed method is also compatible with sliding window-based search using the similar technique developed for MSEA (Wei and Lai, 2007): since MSEA is based on the sum of squared intensities within each block, its computation is efficiently computed even with sliding windows using the well-known integral image technique (Viola and Jones, 2004). The proposed method is based on the sum of intensities and the sum of squared intensities, its computation can be accelerated using two integral images for intensities and squared intensities. This can be yet another benefit of the proposed method over DCT.

6. Conclusion

We proposed new low-dimensional image features for very efficient matching of images. The proposed features can provide an accurate approximation of NCC by a simple dot product between low-dimensional features. This is especially effective with very high recall (more than 95%), and the high accuracy of image matching was demonstrated by actual image sets obtained from broadcast videos. Many advantages of the features were demonstrated compared with other well-know image features, viz., DCT coefficients, histograms, and MSEA.

Although the proposed features were primarily designed for images, the features can be adapted to other types of data with Markovian properties, such as time-series, volume, or spatio-temporal data (viz., video). The current design of features only took into account the correlation between pixel intensities within close proximity, however, if we can employ correlation between successive frames of video, duplicate shots can be detected very efficiently. Images were matched in the current implementation by linear scanning. The combination with index structures such as trees or hashes would probably also be promising directions for further study.

Appendix A. Derivation of upper and lower bounds of NCC

By using the method of Lagrange multipliers, dΛ = 0 results in 2 mn equations ($\partial\Lambda/\partial x_{i,j} = 0$ and $\partial\Lambda/\partial y_{i,j} = 0$ for all i and j). From these equations, we obtain

$$4\lambda_{2,i}\lambda_{4,i} - 1 = 0, \tag{A.1}$$

$$2\lambda_{1,i}\lambda_{4,i} + \lambda_{3,i} = 0, \tag{A.2}$$

$$y_{i,j} = 2\lambda_{2,i}x_{i,j} + \lambda_{1,i}. \tag{A.3}$$

Combining these with Eqs. (8)–(11), we obtain

$$Y_i = \sum_j y_{i,j} = \sum_j \left(2\lambda_{2,i}x_{i,j} + \lambda_{1,i}\right) = 2\lambda_{2,i}X_i + m\lambda_{1,i}, \tag{A.4}$$

$$YY_i = \sum_j y_{i,j}^2 = \sum_j \left(2\lambda_{2,i}x_{i,j} + \lambda_{1,i}\right)^2 = 4\lambda_{2,i}^2 XX_i + 4\lambda_{1,i}\lambda_{2,i}X_i + m\lambda_{1,i}^2. \tag{A.5}$$

Solving these quadratic equations with $2n$ unknowns ($\lambda_{1,i}$, $\lambda_{2,i}$, $i = 1,\ldots,n$), we obtain

$$\lambda_{1,i} = -\frac{2\lambda_{2,i}}{m}X_i + \frac{1}{m}Y_i, \tag{A.6}$$

$$\lambda_{2,i} = \pm\frac{1}{2}\sqrt{\frac{|mYY_i - Y_i^2|}{|mXX_i - X_i^2|}}. \tag{A.7}$$

Finally we combine these with Eq. 7 to obtain NCC

$$\mathrm{NCC}(x,y) = \sum_{i,j} x_{i,j}y_{i,j} = \sum_{i,j} x_{i,j}\left(2\lambda_{2,i}x_{i,j} + \lambda_{1,i}\right) = \frac{1}{m}\sum_i \left(X_iY_i \pm \sqrt{\left(mXX_i - X_i^2\right)\left(mYY_i - Y_i^2\right)}\right). \tag{A.8}$$

The result with the plus sign for the plus-minus sign gives the upper-bound, and the result with the minus sign gives the lower-bound.

Appendix B. Comparison of upper-bounds of the proposed method and MSEA

Let us compare the upper-bound of the proposed method (Eq. (15)) and the upper-bound of MSEA (Eq. (25)) in terms of the tightness. Now Eq. (25) can be converted as

$$UB^{\mathrm{MSEA}} = \sum_i \left(\frac{X_iY_i}{m} + \sqrt{\left(\sqrt{XX_iYY_i} - \frac{X_iY_i}{m}\right)^2}\right). \tag{B.1}$$

Comparing this and Eq. (15), we simply need to see the difference between the second terms of Eqs. (15) and (B.1),

$$\left(XX_i - \frac{X_i^2}{m}\right)\left(YY_i - \frac{Y_i^2}{m}\right) - \left(\sqrt{XX_iYY_i} - \frac{X_iY_i}{m}\right)^2 = -\frac{1}{m}\left(\left(XX_iY_i^2 + X_i^2YY_i\right) - 2X_iY_i\sqrt{XX_iYY_i}\right) = -\frac{1}{m}\left(X_i\sqrt{YY_i} - Y_i\sqrt{XX_i}\right)^2 \leqslant 0. \tag{B.2}$$

This proves that

$$UB \leqslant UB^{\mathrm{MSEA}} \tag{B.3}$$

in other words, the proposed method gives theoretically tighter upper-bound than MSEA. Note that this holds when the same block decomposition for images is used.

References

Agrawal, R., Faloutsos, C., Swami, A., 1993. Efficient similarity search in sequence databases. In: Proc. FODO. Springer-Verlag, pp. 69–84.

Andoni, A., Indyk, P., 2008. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. Comm. ACM 51 (1), 117–122.

Arya, S., Mount, D.M., Netanyahu, N.S., Silverman, R., Wu, A.Y., 1998. An optimal algorithm for approximate nearest neighbor searching fixed dimensions. J. ACM 45, 891–923.

Ballard, D.H., Brown, C.M., 1982. Computer Vision, first ed. Prentice Hall, Professional Technical Reference.

Berrani, S.-A., Manson, G., Lechat, P., 2008. A non-supervised approach for repeated sequence detection in TV broadcast streams. Image Comm. 23, 525–537.

Böhm, C., Berchtold, S., Keim, D.A., 2001. Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. ACM Comput. Surveys 33 (3), 322–373.

Clarke, R., 1981. Relation between the Karhunen–Loève and cosine transforms. Proc. IEE, Part F 128 (6), 359–360.

Di Stefano, L., Mattoccia, S., 2003. Fast template matching using bounded partial correlation. Machine Vision Appl. 13, 213–221.

Döhring, I., Lienhart, R., 2009. Mining tv broadcasts for recurring video sequences. In: Proc. ACM Internat. Conf. on Image and Video Retrieval. CIVR '09, ACM, New York, NY, USA, pp. 28:1–28:8.

Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L., Schmid, C., 2009. Evaluation of gist descriptors for web-scale image search. In: Proc. ACM Internat. Conf. on Image and Video Retrieval. CIVR '09, ACM, New York, NY, USA, pp. 19:1–19:8.

Forsyth, D.A., Ponce, J., 2002. Computer Vision: A Modern Approach. Prentice Hall, Professional Technical Reference.

Gao, X., Duanmu, C.J., Zou, C.R., 2000. A multilevel successive elimination algorithm for block matching motion estimation. IEEE Trans. Image Process. 9 (3), 501–504.

Hirschmüller, H., Scharstein, D., 2009. Evaluation of stereo matching costs on images with radiometric differences. IEEE Trans. Pattern Anal. Machine Intell. 31 (9), 1582–1599.

Hsu, W.H., Chang, S.-F., 2006. Topic tracking across broadcast news videos with visual duplicates and semantic concepts. In: Proc. ICIP.

Jing, Y., Baluja, S., 2008. PageRank for product image search. In: Proc. WWW2008, pp. 307–316.

Joly, A., Buisson, O., Frelicot, C., 2007. Content-based copy detection using distortion-based probabilistic similarity search. IEEE Trans. Multimedia 9 (2), 293–306.

Law-To, J., Buisson, O., Gouet-Brunet, V., Boujemaa, N., 2006. Robust voting algorithm based on labels of behavior for video copy detection. In: ACM Multimedia.

Lejsek, H., Asmundsson, F., Jónsson, B., Amsaleg, L., 2006. Scalability of local descriptors: A comparative study. In: ACM Multimedia.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Internat. J. Comput. Vision 60 (2), 91–110.

Mattoccia, S., Tombari, F., di Stefano, L., 2008. Fast full-search equivalent template matching by enhanced bounded correlation. IEEE Trans. Image Process. 17 (4), 528–538.

Nister, D., Stewenius, H., 2006. Scalable recognition with a vocabulary tree. In: Proc. CVPR, vol. 2, pp. 2161–2168.

Satoh, S., 2004. Generalized histogram: Empirical optimization of low dimensional features for image matching. Proc. European Conf. on Computer Vision (ECCV2004), vol. III. Springer-Verlag, pp. 210–223.

Tsai, D.-M., Lin, C.-T., 2003. Fast normalized cross correlation for defect detection. Pattern Recognition Lett. 24 (15), 2625–2631.

Viola, P., Jones, M., 2004. Robust real-time face detection. Internat. J. Comput. Vision 57 (2), 137–154.

Wang, X.-J., Zhang, L., Liu, M., Li, Y., Ma, W.-Y., 2010. Arista – image search to annotation on billions of web photos. In: CVPR, pp. 2987–2994.

Wei, S.-D., Lai, S.-H., 2007. Efficient normalized cross correlation based on adaptive multilevel successive elimination. In: Proc. 8th Asian Conf. on Computer Vision – Volume Part I ACCV'07. Springer-Verlag, Berlin, Heidelberg, pp. 638–646.

Wu, X., Ngo, C.-W., Li, Q., 2006. Threading and autodocumenting news videos. IEEE Signal Process. Magazine 23 (2), 59–68.

Wu, X., Putpuek, N., Satoh, S., 2010. Commercial film detection and identification based on a dual-stage temporal recurrence hashing algorithm. In: Workshop on Very-Large-Scale Multimedia Corpus, Mining and Retrieval (VLSMCMR2010), in Conjunction with ACM Multimedia, 2010.

Yamagishi, F., Satoh, S., Hamada, T., Sakauchi, M., 2003. Identical video segment detection for large-scale broadcast video archives. In: Proc. Internat. Workshop on Content-Based Multimedia Indexing (CBMI'03), pp. 135–142.

Yamagishi, F., Satoh, S., Sakauchi, M., 2004. A news video browser using identical video segment detection. In: Proc. Pacific-Rim Conf. on Multimedia (PCM2004), vol. II, pp. 205–212.

Zhai, Y., Shah, M., 2005. Tracking news stories across different sources. In: ACM Multimedia.

Zhao, W., Ngo, C.-W., Tan, H.-K., Wu, X., 2007. Near-duplicate keyframe identification with interest point matching and pattern learning. IEEE Trans. Multimedia 9 (5), 1037–1048.