Signal Processing: *Image Communication* 26 (2011) 612–627

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

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

Bimodal fusion of low-level visual features and high-level semantic features for near-duplicate video clip detection

Hyun-seok Min, Jae Young Choi, Wesley De Neve, Yong Man Ro *

Image and Video Systems Lab, Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 305-732, Republic of Korea

ARTICLE INFO

Article history:
Received 14 October 2010
Accepted 2 April 2011
Available online 15 April 2011

Keywords:
NDVC detection
Near-duplicates
Semantic features
Video copy detection
Video signatures

ABSTRACT

The detection of near-duplicate video clips (NDVCs) is an area of current research interest and intense development. Most NDVC detection methods represent video clips with a unique set of low-level visual features, typically describing color or texture information. However, low-level visual features are sensitive to transformations of the video content. Given the observation that transformations tend to preserve the semantic information conveyed by the video content, we propose a novel approach for identifying NDVCs, making use of both low-level visual features (this is, MPEG-7 visual features) and high-level semantic features (this is, 32 semantic concepts detected using trained classifiers). Experimental results obtained for the publicly available MUSCLE-VCD-2007 and TRECVID 2008 video sets show that bimodal fusion of visual and semantic features facilitates robust NDVC detection. In particular, the proposed method is able to identify NDVCs with a low missed detection rate (3% on average) and a low false alarm rate (2% on average). In addition, the combined use of visual and semantic features outperforms the separate use of either of them in terms of NDVC detection effectiveness. Further, we demonstrate that the effectiveness of the proposed method is on par with or better than the effectiveness of three state-of-the-art NDVC detection methods either making use of temporal ordinal measurement, features computed using the Scale-Invariant Feature Transform (SIFT), or bag-of-visual-words (BoVW). We also show that the influence of the effectiveness of semantic concept detection on the effectiveness of NDVC detection is limited, as long as the mean average precision (MAP) of the semantic concept detectors used is higher than 0.3. Finally, we illustrate that the computational complexity of our NDVC detection method is competitive with the computational complexity of the three aforementioned NDVC detection methods.

1. Introduction

Easy-to-use multimedia devices and cheap storage and bandwidth enable users to upload a vast amount of digital video content to websites for video sharing (e.g., YouTube and Vimeo). Since digital video content can be easily edited and redistributed, websites for video sharing contain a high number of similar video clips. These video clips are usually referred to as duplicate video clips (exact copies) or near-duplicate video clips (NDVCs or transformed copies) [1,2].

Websites for video sharing have a strong need for detecting NDVCs. The detection of NDVCs is for instance an important prerequisite for the protection of intellectual property (IP) [3,4]. In addition, the ability to detect NDVCs makes it possible to mitigate visual redundancy

* Corresponding author. Tel.: +82 42 350 3494;
fax: +82 42 350 6245.
E-mail addresses: hsmin@kaist.ac.kr (H.-s. Min), jygchoi@kaist.ac.kr (J.Y. Choi), wesley.deneve@kaist.ac.kr (W. De Neve), ymro@ee.kaist.ac.kr (Y.M. Ro).

0923-5965/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2011.04.001

when presenting video search results (e.g., by clustering duplicates and near-duplicates) [2]. Further applications of NDVC detection include media usage monitoring, content linking on the Web, metadata propagation for annotation purposes, and management of personal media libraries [5].

When building a system for detecting NDVCs, it is important to have a working definition of NDVCs. At the time of writing, no agreement exists on the technical definition of an NDVC, due to the ambiguity of the term "near". As such, definitions differ, depending on the transformations allowed and the applications targeted [1]. Most definitions characterize NDVCs as approximately identical video clips that have been the subject of one or more transformations. The aforementioned definition is most useful when targeting the detection of copyright infringement. Possible alterations of the video content include photometric transformations (e.g., change of color and lighting), editing operations (e.g., insertion of captions, logos, and borders), and spatiotemporal transformations (e.g., change of resolution and frame rate) [2].

A significant number of NDVC detection methods represent video clips with a unique set of low-level visual features, a representation commonly referred to as a video signature or video fingerprint. The low-level visual features, typically extracted from keyframes, may for instance describe color [3] or the spatial distribution of intensity values [6,7]. Video signatures need to be robust with respect to (significant) transformations of the video content. However, video signatures using low-level visual features are often sensitive to content transformations [6]. As such, video signatures using low-level visual features do not perform well for detecting NDVCs.

Although content transformations may significantly modify low-level visual features, the transformations applied tend to preserve the semantic information conveyed by the original video content [1,8]. Therefore, in this paper, we investigate the use of semantic features for the purpose of NDVC detection. This research direction is also in line with the conclusions of [1], presenting an extensive study on the user perception of NDVCs. In particular, one conclusion of [1] is that the detection of semantic similarity can likely improve the effectiveness of NDVC detection.

Although semantic concept detection can be considered a long-time issue in the field of multimedia retrieval [7,9], to the best of our knowledge, a semantic approach towards the detection of NDVCs has not been thoroughly investigated yet. In this context, two important challenges need to be taken into account, both related to the Semantic Gap [10,11]:

- *Semantic coverage:* Semantic concept detection is typically realized using a finite number of trained classifiers. It should be clear that the discriminative power of a video signature can be considered low when this video signature only represents the global presence or absence of a limited number of semantic concepts in a video clip [12].
- *Effectiveness of semantic concept detection*: Despite substantial progress made during the past years [13], the effectiveness of trained classifiers is still limited, leaving room for significant improvement [14].It should be clear that the effectiveness of semantic concept detection might affect the effectiveness of a semantic approach towards the problem of NDVC detection.

To overcome a limited semantic coverage caused by the use of a restricted semantic concept vocabulary, we take advantage of the *variation* of several popular semantic concepts along the temporal axis of a video clip. Indeed, although video clips may contain similar semantic concepts from a global perspective, we assume that the temporal variation of semantic concepts is different from video clip to video clip. To minimize the influence of the limited effectiveness of semantic concept detection on the effectiveness of NDVC detection, we propose to make use of trained classifiers that come with reasonably high detection effectiveness [13].

Further, to improve the discriminative power of the proposed NDVC detection method, we introduce a *hybrid video matching method* that makes use of both low-level visual features and high-level semantic features (i.e., high-level semantic concepts). The use of low-level visual features allows handling a low discriminative power when the temporal variation of the semantic concepts used is low. As such, bimodal fusion of visual and semantic features allows combining their strengths and eliminating their weaknesses.

The proposed NDVC detection method has been evaluated using two publicly available video sets: TRECVID 2008 [15] and MUSCLE-VCD-2007 [16]. Experimental results show that our hybrid approach is able to identify NDVCs with a low missed detection probability (3% on average) and a low false alarm rate (2% on average). In addition, the combined use of visual and semantic features outperforms the separate use of either of them in terms of NDVC detection effectiveness. Further, we demonstrate that the effectiveness of the proposed method is on par with or better than the effectiveness of three state-of-the-art NDVC detection methods either making use of temporal ordinal measurement, features computed using the Scale-Invariant Feature Transform (SIFT), or bag-of-visual-words (BoVW). We also show that the influence of the effectiveness of semantic concept detection on the effectiveness of NDVC detection is limited, as long as the mean average precision (MAP) of the semantic concept detectors is higher than 0.3. Finally, we illustrate that the computational complexity of our NDVC detection method is competitive with the computational complexity of the three aforementioned NDVC detection methods.

The rest of the paper is organized as follows. Section 2 provides an overview of related work. We pay particular attention to different NDVC definitions that are currently in use, arguing that our semantic approach towards the task of NDVC detection is meaningful irrespective of the NDVC definition used. Section 3 briefly discusses the creation of a semantic video signature. This section also explains our hybrid video matching method. Section 4 describes our evaluation methodology, whereas Section 5 presents and discusses our experimental results. Finally, Section 6 outlines conclusions and directions for future research.

Note that this paper is an extended and improved version of conference submissions [17,18], offering an integrated and more rigorous treatment of the proposed method for NDVC detection. In particular, in this paper, we propose a hybrid approach towards the task of NDVC detection, making use of both visual and semantic features in order to characterize video clips in a robust way. In addition, using the MUSCLE-VCD-2007 and the TRECVID 2008 video sets, we present experimental results that are more extensive, studying the influence of semantic coverage and the effectiveness of semantic concept detection on the effectiveness of NDVC detection.

2. Related work

In this section, we review related work. We pay particular attention to different NDVC definitions that are currently in use, arguing that our semantic approach towards the task of NDVC detection is meaningful irrespective of the NDVC definition used. Similar to [2,5], we consider content-based methods for NDVC detection as complementary to text- and context-based approaches.

2.1. NDVC definitions

As mentioned in the introduction of this paper, the scientific literature contains several NDVC definitions, due to the ambiguity of the term “near”. The basic definition of an NDVC can be summarized as follows: *NDVCs are approximately identical video clips that have been the subject of at least one transformation, and where the transformations applied preserve the semantic information conveyed by the original video clip* [2,9]. Recently, a number of papers have extended the basic definition of an NDVC in order to include semantic aspects. For example, the authors of [19,20] indicate that the definition of an NDVC should cover video clips of the same scene (e.g., a person riding a bike), and where these video clips capture the scene in question using different viewpoints. Similarly, the user study presented in [1] indicates that the definition of an NDVC should also include a user-centric component. This is motivated by the following two observations: (1) identical video clips containing relevant complementary information are not considered as NDVCs by users and (2) video clips that are not alike, but that are visually similar and semantically related are also perceived as NDVCs by users. As for the latter observation, the same semantic concepts must be present without relevant additional information (i.e., the same information is presented under different scene settings).

The basic NDVC definition is useful for applications that aim at detecting copyright infringement, media usage monitoring, and management of personal media libraries, whereas the extended NDVC definition is of particular interest when targeting visual redundancy elimination in video search results, content linking on the Web, and metadata propagation. It should be clear that our semantic approach towards the task of NDVC detection is relevant to both the basic and extended definition of an NDVC.

2.2. NDVC detection based on low-level visual features

The past few years have witnessed the development of several content-based methods for NDVC detection. These methods typically focus on the creation of video signatures that make use of low-level visual features. When taking into account the type of visual features used, video signatures for content-based NDVC detection can be divided into the following four categories: global, local, spatial, and temporal signatures. This is illustrated by Table 1.

2.2.1. Global and local signatures

Images or video frames can be described using either a global signature or a set of local signatures [9]. Global signatures describe the whole image region in terms of characteristics such as color (e.g., color histogram [3]), texture, or the spatial distribution of intensity values (e.g., ordinal signature [7]). The creation of global signatures can be considered straightforward. In addition, global signatures work well for detecting NDVCs with slightly altered low-level visual features. However, global signatures do not perform well for NDVCs with significantly altered low-level visual features. For example, the discriminative power of video signatures using global signatures significantly decreases when applying transformations such as cropping and zooming (despite the fact that cropping and zooming only partly modify video frames) [26].

The idea behind the use of local signatures is to detect and describe small regions-of-interest or patches in a video frame. Similarity between different frames can then be determined by counting the number of similar patches the frames have in common. Local signatures are highly robust to geometric and photometric variations [2].

Table 1
Different types of content-based video signatures.

	Global signatures (frame level)	Local signatures (patch level)
Spatial signatures (image signatures)	Color histogram [3], ordinal intensity [7], spatial correlation [21,22]	Scale-Invariant Feature Transform (SIFT) [27], Speeded Up Robust Features (SURF) [29], Local Binary Pattern (LBP) [29], Local Difference Pattern (LDP) [28]
Temporal signatures (video signatures)	Temporal ordinal measure [23], motion [24], tomography [25], camera motion [26]	Trajectory of interest points [30], shot-based interest points [31]

However, their robustness remains unclear when a video clip is suffering from a low spatial resolution, motion blur, or compression artifacts [30]. Furthermore, the number of interest points that needs to be computed for each frame is significant, resulting in a high computational complexity [2].

2.2.2. Spatial and temporal signatures

Images or video frames can also be described using spatial or temporal signatures. Spatial signatures refer to signatures that only make use of features from a single frame for their construction. This is in contrast to temporal signatures, which make use of multiple frames for their construction. In particular, temporal signatures typically exploit the variation in object or camera motion between several frames. The temporal variation in object or camera motion is less affected by spatial transformations such as the addition of Gaussian noise or a shift in illumination. However, the robustness of temporal signatures may suffer when a video clip has been the subject of temporal transformations.

2.3. NDVC detection based on high-level semantic features

The authors of [8] make use of the appearance and the identity of human faces in order to characterize a video segment, targeting fingerprinting and retrieval in large-scale databases containing motion picture and television data. Specifically, pulse-like signals that provide information on whether a human face appears or not in each frame of a video clip are used to construct a video signature. As argued in [8], the use of semantic information makes the proposed video signature highly robust to video noise and other types of video processing. However, as acknowledged by the authors of [8], by only taking advantage of face information, the proposed approach is less suitable for characterizing sports content, news content, and documentaries. Similar to [8], the authors of [32] make use of the appearance of human faces in order to realize NDVC detection. In particular, the authors of [32] fuse the results of three different techniques for identifying NDVCs: shot matching using face and extended body information, activity subsequence matching, and non-facial shot matching using low-level visual feature similarity.

2.4. NDVC detection using hybrid approaches

Fusing different types of features is a common way to improve the effectiveness of NDVC detection. For example, the authors of [33] propose a semi-global descriptor, combining local and global information. Specifically, local information is taken into account by partitioning a frame into four regions, subsequently characterizing each region by means of the global MSF-Color descriptor (MSF is short for Markov Stationary Feature). The authors of [28] identify NDVCs using a hierarchical approach. An initial attempt to quickly detect NDVCs is realized using signatures derived from global color histograms. When global features cannot be used to unambiguously identify a query video clip as either novel or as an NDVC, local features are used, allowing for a more accurate analysis at the expense of additional computational complexity. [34] discusses a so-called glocal signature for the purpose of NDVC detection, extending the description of local keypoints with information regarding the geometric configuration of these keypoints in the image plane.

Recently, a number of research efforts have been presented that fuse both content and context information to enhance the identification of NDVCs. [35] reports that the effectiveness of retrieving near-duplicate keyframes can be improved by linearly fusing visual keywords and semantic context derived from the speech script surrounding a keyframe. Similarly, the method proposed in [5] combines visual features and context information (e.g., title, time duration, description, and view count) for detecting NDVCs.

3. Proposed method for NDVC detection

Given a query video clip, Fig. 1 shows that the proposed NDVC detection method largely consists of five sequential steps. These steps are as follows: (1) extraction of low-level visual features for each shot in the query video clip; (2) creation of a semantic vector for each shot; (3) creation of a semantic video signature; (4) measurement of temporal semantic variation; and (5) hybrid video matching.

After shot segmentation, keyframe selection, and visual feature extraction, we perform semantic concept detection for each shot, relying on trained classifiers that come with reasonably high detection effectiveness. That way, we are able to represent each shot by a *semantic vector*, denoting the presence or absence of the semantic concepts used in the shot under consideration. By aggregating the semantic vectors of all shots in the query video clip, we are then able to create a *semantic video signature* for the query video clip. Using the semantic video signature constructed for the query video clip, we subsequently compute the variation of the semantic features along the temporal dimension of the query video clip. Next, to determine whether the query video clip is an NDVC, video matching measures the similarity between the query video clip and the reference video clips stored in the reference video database, making use of both visual and semantic features.

Shot segmentation, keyframe selection, and visual feature extraction are described in more detail in Section 4, as part of our experimental setup, whereas the creation of a semantic video signature, the measurement of temporal semantic variation, and the hybrid matching of video signatures are discussed in the following subsections.

3.1. Creation of a semantic video signature using trained classifiers

Video content has an outspoken temporal structure: video clips consist of scenes; scenes consist of shots; and shots consist of frames. Video shots are widely used as the fundamental unit of processing in the field of video content analysis [36]. Therefore, in order to extract

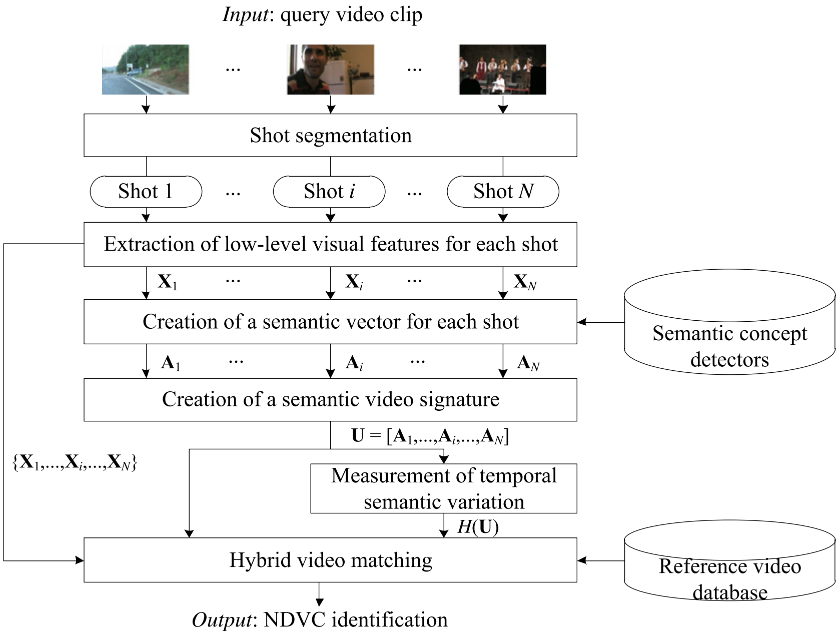
Input: query video clip
Shot segmentation
Shot 1
Shot i
Shot N
Extraction of low-level visual features for each shot
X1 … Xi … XN
Creation of a semantic vector for each shot
Semantic concept detectors
A1 … Ai … AN
Creation of a semantic video signature
U = [A1,...,Ai,...,AN]
{X1,...,Xi,...,XN}
Measurement of temporal semantic variation
H(U)
Hybrid video matching
Reference video database
Output: NDVC identification

Fig. 1. Overview of the proposed method for bimodal NDVC detection using visual and semantic features.

semantic concepts, a video clip **V** is first segmented into N shots such that $\mathbf{V}=\{\mathbf{S}_i\}_{i=1}^{N}$, where $\mathbf{S}_i$ denotes the ith shot of **V**. The set of low-level visual features of $\mathbf{S}_i$ is then denoted as $\mathbf{X}_i=\{\mathbf{x}_{i,j}\}_{j=1}^{R}$, where $\mathbf{x}_{i,j}$ represents the jth low-level visual feature extracted from $\mathbf{S}_i$ and where R denotes the total number of low-level visual features extracted.

Let $\mathbf{C}=\{c_k\}_{k=1}^{M}$be a set consisting of M different semantic concepts, with c_k representing the kth semantic concept (e.g., ‘sky’, ‘indoor’, ‘face’, or ‘architecture’). Also, let $d_k(\cdot)$ denote a classifier function that is used for detecting semantic concept c_k. To this end, given the set of low-level visual features $\mathbf{X}_i$ extracted from $\mathbf{S}_i$, the classifier function $d_k(\cdot)$ produces a confidence value for the kth semantic concept c_k. It is common to make use of the a posterior probability $P(c_k|\mathbf{X}_i)$ to represent the conditional probability that the semantic concept c_k is present within a shot $\mathbf{S}_i$ represented by the low-level visual features $\mathbf{X}_i$ [37]. Using $P(c_k|\mathbf{X}_i)$, we determine whether concept c_k is absent or present in shot $\mathbf{S}_i$ as follows:

$$a_{i,k}=\begin{cases}1 & \text{if } P(c_k|\mathbf{X}_i)\geq \xi_k \\ 0 & \text{otherwise}\end{cases}, \tag{1}$$

where ζ_k is a pre-specified threshold for c_k. In (1), if $P(c_k|\mathbf{X}_i)$ is higher than the threshold ζ_k, we decide that shot $\mathbf{S}_i$ contains c_k, in which case $a_{i,k}$ becomes one. Otherwise, we decide that shot $\mathbf{S}_i$ does not contain c_k, in which case $a_{i,k}$ becomes zero. Note that ζ_k can be obtained using the heuristic method proposed in [37]. Also, note that any classifier can be used to construct the classifier function $d_k(\cdot)$ (e.g., Support Vector Machines (SVMs)).

A single shot usually contains multiple semantic concepts. To account for this observation, we define the *semantic vector* $\mathbf{A}_i=[a_{i,1},\ a_{i,2},\ldots,a_{i,M}]^{\mathrm{T}}$ for $\mathbf{S}_i$, where T is the transpose operator. $\mathbf{A}_i$ denotes the presence or absence in $\mathbf{S}_i$ of the M semantic concepts used. Using $\mathbf{A}_i$, we then define the *semantic signature* of **V** as follows:

$$\mathbf{U}=[\mathbf{A}_1\quad \mathbf{A}_2\cdots\mathbf{A}_N], \tag{2}$$

In (2), **U** is a matrix of dimension M by N, containing binary values. As shown in Fig. 2, **U** captures the temporal variation of M different semantic concepts in N shots of **V**.

3.2. *Measuring temporal semantic variation using temporal entropy*

A semantic video signature has limited discriminative power when it only relies on the global presence or absence of a limited number of semantic concepts for its construction. In order to increase the discriminative power of a semantic video signature, we propose to take advantage of the variation of semantic features along the temporal axis of a video clip. If the temporal variation of the semantic features is high, then the discriminative power of the proposed semantic video signature will also be high. However, if the temporal variation of the semantic features is low, then the discriminative power of the proposed semantic video signature will still be low.

Fig. 3 shows two (special) cases in which the temporal variation of the semantic features is low. First, as illustrated by Fig. 3(a), the semantic concepts used may not (completely) cover the semantic information conveyed by the video clip under consideration (in this case, the semantic concepts ‘moon’ and ‘satellite’ are not part of the semantic concept vocabulary used). Second, as illustrated by Fig. 3(b), the video clip under consideration may only contain a small number of semantic concepts. As a result, in both of the aforementioned cases, most elements in the semantic vectors are equal to zero. This is also illustrated by the two histograms presented in Fig. 3(a) and (b), showing that the distribution of the

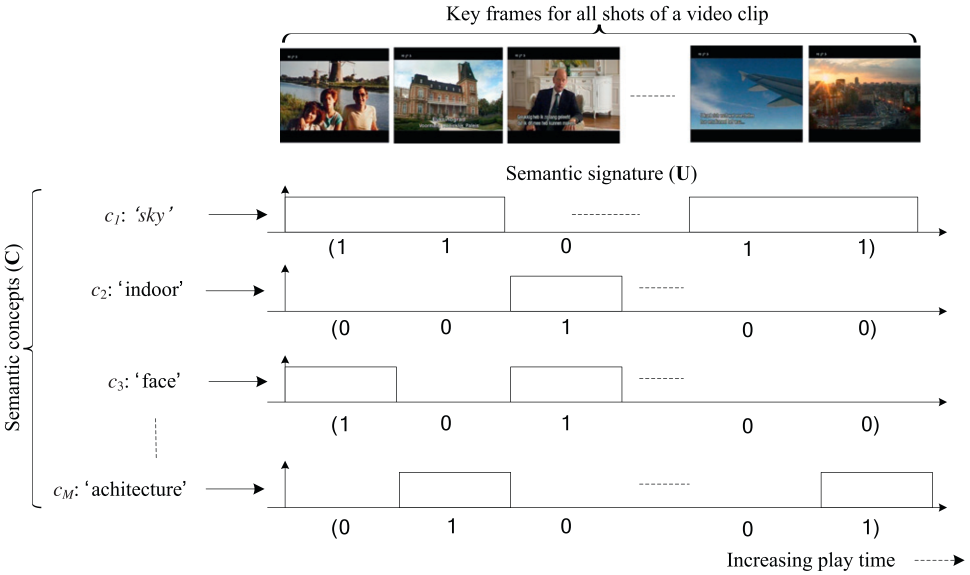

Fig. 2. Creation of a semantic video signature. As described by (1), the absence and presence of a particular semantic concept (c_k) is indicated by a zero and a one, respectively, (e.g., '101…00' for the semantic concept 'face').

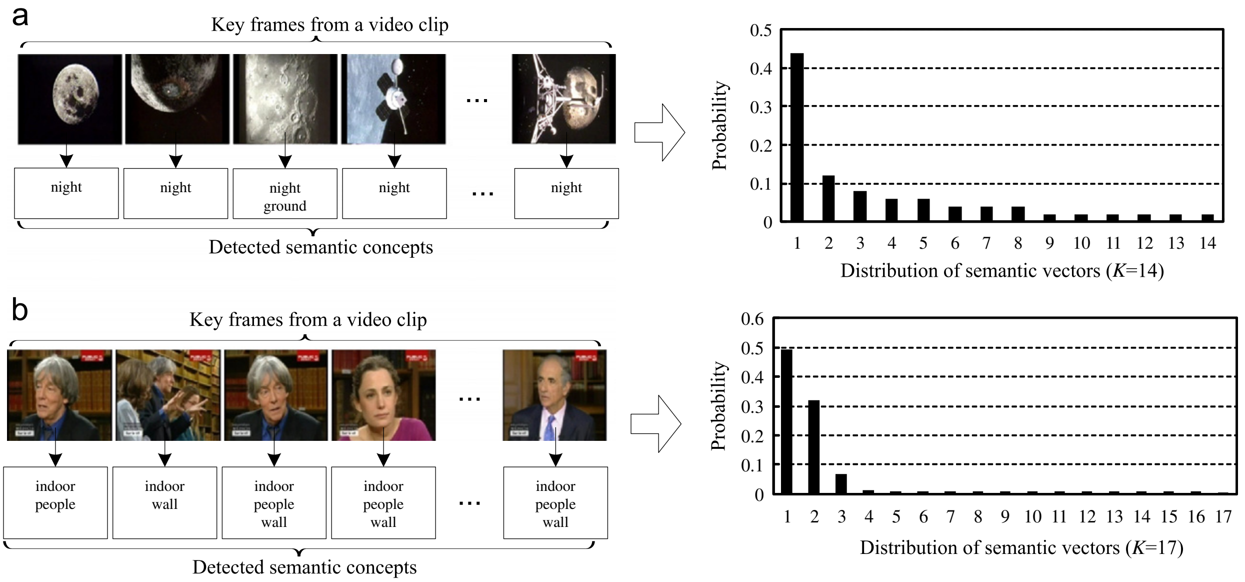

Fig. 3. Low temporal semantic variation: (a) limited semantic coverage by the concept vocabulary used and (b) a query video clip containing a small number of semantic concepts. The histograms on the right-hand side show the distribution of the semantic vectors for the video clips used.

semantic vectors is highly uneven in both of the aforementioned cases. In other words, only a few semantic vectors are highly popular.

The observations outlined above make clear that we need a mechanism to determine the temporal variation of the semantic features used. In addition, when detecting that the semantic features have low temporal variation, we need a mechanism to maintain a feasible amount of discriminative power. As discussed in more detail in Section 3.3, we address the latter need by also making use of low-level visual features to identify NDVCs. We address the former need by computing the temporal entropy of the semantic features used. High values of temporal entropy indicate high semantic variation and thus a semantic modality with high discriminative power, whereas low values of temporal entropy signal low semantic variation and thus a semantic modality with low discriminative power.

To determine the temporal entropy $H(\mathbf{U})$ of the semantic video signature $\mathbf{U}$, let us assume that K semantic vectors out of a total of N semantic vectors are different within $\mathbf{U}$, with $K \le N$. We compute the probability P_i that a semantic vector $\mathbf{A}_i$ comes with a particular temporal variation $\mathbf{A}_j$ as follows:

$$P_i = \frac{\sum_{j=1}^{N} \delta_i(\mathbf{A}_j)}{N}, \quad \text{for} \quad i = 1, \ldots, K, \tag{3}$$

where $\delta_i(\cdot)$ is a function that returns one if a match occurs between $\mathbf{A}_i$ and $\mathbf{A}_j$, and zero otherwise. Using (3), the temporal entropy $H(\mathbf{U})$ of the semantic video signature $\mathbf{U}$ can then be computed as follows:

$$H(\mathbf{U}) = -\alpha \sum_{i=1}^{K} P_i \log_2 P_i, \tag{4}$$

Note that α is a normalizing factor, denoting the maximum entropy value. This maximum is achieved when all N semantic vectors are different. As such, α is computed as $1/\log_2 N$, where N represents the total number of shots. Note that $H(\mathbf{U})$ becomes zero when all N semantic vectors are identical. In other words, if the temporal variation of the semantic features in $\mathbf{V}$ is low, then $H(\mathbf{U})$ will be close to zero. Likewise, if the temporal variation of the semantic features in $\mathbf{V}$ is high, then $H(\mathbf{U})$ will be close to one.

3.3. Hybrid video matching using visual and semantic features

As explained in Section 3.2, low values of temporal entropy signal low semantic variation and thus a semantic modality with low discriminative power. To overcome a semantic modality with low discriminative power, we introduce a hybrid video matching strategy in this section, simultaneously fusing visual features and semantic features. As explained further in this section, fusion is done by again taking advantage of the temporal entropy of the semantic features used, allowing to automatically determine an appropriate weight for the semantic and visual features used in our NDVC detection method.[1]

The objective of video matching is to determine whether a query video clip appears in a reference video clip, and if so, at what location in the reference video clip. Let us denote the query and reference video clips as $\mathbf{V}^q = \{\mathbf{S}_i^q\}_{i=1}^N$ and $\mathbf{V}^t = \{\mathbf{S}_l^t\}_{l=1}^L$, where $\mathbf{S}_i{}^q$ and $\mathbf{S}_l{}^t$ represent the ith and lth shot of $\mathbf{V}^q$ and $\mathbf{V}^t$, respectively. It is reasonable to assume that $N < L$. Indeed, video copies usually contain only a portion of the reference video clip (rather than containing the whole reference video clip). This can be attributed to technical reasons (e.g., upload limits) or content reasons (e.g., sharing of movie highlights) [39].

Using (2), we denote the semantic video signatures of $\mathbf{V}^q$ and $\mathbf{V}^t$ as $\mathbf{U}^q = [\mathbf{A}_1^q \quad \mathbf{A}_2^q \cdots \mathbf{A}_N^q]$ and $\mathbf{U}^t = [\mathbf{A}_1^t \quad \mathbf{A}_2^t \cdots \mathbf{A}_L^t]$, respectively. Note that each element of the semantic vectors $\mathbf{A}_i{}^q$ $(i=1, \ldots, N)$ and $\mathbf{A}_l{}^t$ $(i=1, \ldots, L)$ can be obtained by making use of (1). In addition, we denote the set of low-level visual features of the shots $\mathbf{S}_i{}^q$ and $\mathbf{S}_l{}^t$ as $\mathbf{X}_i{}^q = (\mathbf{x}_{i,j}^q)_{j=1}^R$ and $\mathbf{X}_l{}^t = (\mathbf{x}_{i,j}^t)_{j=1}^R$, respectively, with R denoting the total number of ow-level visual features used. The dissimilarity between $\mathbf{V}^q$ and $\mathbf{V}^t$ is then computed as follows:

$$d(\mathbf{V}^q, \mathbf{V}^t) = \min_p \frac{1}{N} \sum_{i=1}^{N} d_{\mathrm{shot}}(\mathbf{A}_i^q, \mathbf{A}_{i+p}^t, \mathbf{X}_i^q, \mathbf{X}_{i+p}^t), \tag{5}$$

where

$$d_{shot}(\mathbf{A}_i^q, \mathbf{A}_l^t, \mathbf{X}_i^q, \mathbf{X}_l^t) = H(\mathbf{U}^q) \frac{1}{M} \sum_{k=1}^{M} |a_{i,k}^q - a_{l,k}^t| + (1 - H(\mathbf{U}^q)) \frac{1}{R} \sum_{j=1}^{R} ||\mathbf{x}_{i,j}^q - \mathbf{x}_{l,j}^t||^2, \tag{6}$$

and where p denotes the initial shot index (or position) in the reference video clip at which dissimilarity measurement begins, $a_{i,k}^q$ and $a_{i,k}^t$ denote elements (binary values) of the semantic vectors $\mathbf{A}_i{}^q$ and $\mathbf{A}_l{}^t$ respectively, M denotes the number of semantic concepts used, $|\cdot|$ and $\|\cdot\|$ denote the absolute value and norm operations, respectively, and $0 \le H(\mathbf{U}^q) \le 1$. Since $\mathbf{V}^q$ and $\mathbf{V}^t$ usually contain a different number of shots, matching video clips is done by shifting the target video clip $\mathbf{V}^t$ at the level of shots (using a sliding window approach).

In (6), as already indicated in the introduction of this section, the temporal entropy $H(\mathbf{U}^q)$ determines the value of a weighting parameter. This weighting parameter allows for a trade-off between the influence of visual features and the influence of semantic features on the computation of the dissimilarity between $\mathbf{S}_i{}^q$ and $\mathbf{S}_l{}^t$ The higher the discriminative power of the semantic video signature, the higher the influence of the semantic video signature on the dissimilarity computation. On the other hand, when the discriminative power of the semantic video signature is low (i.e., when the value of $H(\mathbf{U}^q)$ is low), the loss in discriminative power of the semantic features is compensated by assigning a higher importance to the influence of the visual features. Consequently, as shown by the experimental results presented in Section 5, our hybrid video matching strategy facilitates robust NDVC detection.

Finally, to determine whether $\mathbf{V}^q$ is a near-duplicate of $\mathbf{V}^t$ or not, we compare the value of $d(\mathbf{V}^q, \mathbf{V}^t)$ with a predetermined threshold ζ_{video}. Specifically, $\mathbf{V}^q$ is considered a near-duplicate of $\mathbf{V}^t$ when the value of $d(\mathbf{V}^q, \mathbf{V}^t)$ is smaller than the value of ζ_{video}. Considering the effectiveness of NDVC identification in our experiments, we found a good compromise by setting ζ_{video} in the range of [0.1, 0.4].

Algorithm 1 summarizes our bimodal approach towards NDVC detection.

Algorithm 1. Bimodal approach for NDVC detection.

```
input: $\mathbf{V}^q$ (query video clip), $\{\mathbf{V}^t\}_{t=1}^T$ (reference video database), $\xi_{video}$
output: $Z$ (decision whether $\mathbf{V}^q$ is an NDVC or not)
Segment a given query video clip $\mathbf{V}^q$ into $N$ shots: $\mathbf{V}^q = \{\mathbf{S}_i^q\}_{i=1}^N$
Extract $\mathbf{X}_i{}^q$ (low-level visual features of $\mathbf{V}^q$) from each shot
  $\mathbf{S}_i^q = \{i=1, \ldots, N\}$
Create $\mathbf{U}^q$ (semantic video signature of $\mathbf{V}^q$) using Eqs. (1) and (2)
for $t := 1$ to $t := T$ do
    comment: $p$ denotes the shot position in $\mathbf{V}^t$ at which similarity
    measurement starts
    for $p := 0$ to $p := L - N$ do
      comment: the distance between $\mathbf{V}^q$ and $\mathbf{V}^t$ is computed
      using Eq. (5)
      $d_{lowest} : = \min(d_{lowest}, d_{video}(\mathbf{V}^q, \mathbf{V}^t))$
    end for
end for
return if $(d_{lowest} < \xi_{video})$ then $Z$: true else $Z$: false.
```

[1] The authors of [38] also make use of entropy to automatically determine the discriminative power of the visual and audio modality used.

4. Experimental setup

4.1. Construction of a reference video database and query video clips

To evaluate the performance of the proposed method for NDVC detection, we conducted an experimental study using two video sets: (1) MUSCLE-VCD-2007, which is a publicly available video set used during the video copy detection evaluation session at the CIVR 2007 conference [16], and (2) the sound and vision data collection used for the search and feature detection effort of TRECVID 2008 [15]. MUSCLE-VCD-2007 contains 101 video clips, with a total duration of 80 h [46]. TRECVID 2008 contains 210 video clips, with a total duration of 90 h. We thus performed experiments with 311 video clips, worth about 170 h of video content. All video clips in the two video sets have a resolution of 352 × 288 pixels and were encoded using MPEG-1. Shot boundary detection was performed using the algorithm proposed in [40]. The rationale behind this choice is threefold [45]: (1) this algorithm, was used to provide the master shot reference for the TRECVID 2008 database; (2) this algorithm is highly effective at detecting hard cuts and various gradual shot changes such as dissolves, fades, and wipes; and (3) this algorithm has a low computational complexity, only using sub-sampled luminance images as input. For the video sets used, we found that the average duration of a shot is 7.74 s. Following the TRECVID 2008 guidelines [41], the frame in the middle of each shot was used as a representative keyframe.

The MUSCLE-VCD-2007 and TRECVID 2008 video sets were used to create a single reference video database. We then randomly selected 100 video clips from the unified reference video database, having a total duration of about 15 h (the average duration of a selected video clip is thus about nine minutes). Query video clips were generated by applying several transformations to the original video clips in the reference video database, focusing on the so-called T6 and T8 transformations used by the content-based video copy detection task of TRECVID 2008 [47,48]. Specifically, the following five transformations were applied to the 100 video clips, resulting in 500 query video clips [31]:

- Blur: frames are blurred using a Gaussian kernel with a radius of 15.
- Crop: frames are cropped to 50% of their size, not preserving the center region.
- Pattern insertion: insertion of a picture with a size 30% of the size of the frame.
- Change in brightness: brightness is increased with 40%.
- Caption insertion: a caption is inserted at the bottom of the frame.

Table 2 summarizes the different types of transformations used in our experiments, as well as a number of relevant input parameters. In addition, Fig. 4 shows the visual effect of the different transformations used.

Table 2
Different types of transformations and parameters used in our experiments.

Transformation	Parameter
Blur	Filter size
Picture-in-picture (PIP)	Size of inserted picture
Subtitles (insertion of captions)	Scale of caption
Cropping	Scale
Change in brightness	Change of brightness

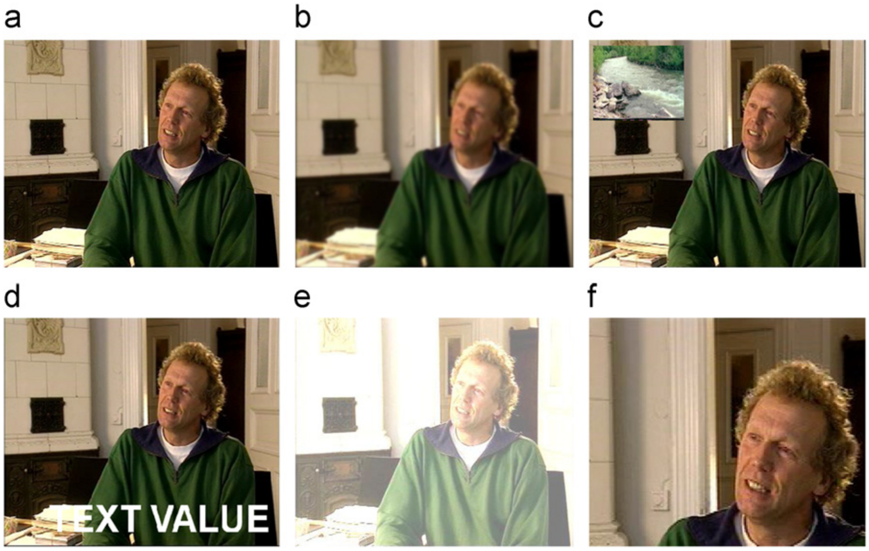

Fig. 4. Example keyframes, showing the visual effect of the different transformations used: (a) original, (b) blur, (c) picture-in-picture, (d) subtitles, (e) change in brightness, and (f) cropping.

4.2. Measurement of NDVC detection effectiveness

The effectiveness of NDVC detection was measured using the ST1 score [16], which is the ratio of the number of correct answers to the total number of query video clips, and the Normalized Detection Cost Ratio (NDCR) [49]. The definition of NDCR is as follows:

$$NDCR = P_{miss} + \beta \times R_{FA}, \tag{7}$$

where

$$P_{miss} = \frac{N_{FN}}{N_{TP}}, \quad R_{FA} = \frac{N_{FP}}{T_{refdata} \times T_{query}}, \tag{8}$$

and where P_{miss} is the probability of a miss and R_{FA} is the false alarm rate. Further, N_{TP}, N_{FN}, and N_{FP} denote the number of true positives, false negatives, and false positives, respectively, while $T_{refdata}$ and T_{query} represent the total duration of the entire reference video database and the query video clips (with duration expressed in hours). In addition, β is a factor that trades off the cost of missing a true positive against the cost of having to deal with a false alarm. In our experiments, following the recommendation made by [49], we set $\beta = 2$.

Note that the higher the ST1 score, the better the effectiveness of NDVC detection (an ST1 score of one indicates perfect results for the transformation considered), whereas the lower the NDCR value, the better the effectiveness of NDVC detection (an NDCR value of zero indicates perfect results for the transformation considered).

4.3. Selection and detection of semantic concepts

Three requirements were used to guide the selection of semantic concepts. First, the selected semantic concepts need to represent visual concepts. Second, the selected semantic concepts need to represent popular concepts, a requirement that can be fulfilled by relying on semantic concepts that describe background information. Third, common machine learning techniques should be able to detect the semantic concepts used in an easy and reliable way. As a result, the following 32 concepts were used in our experiments [17,18]: 'gravel', 'park', 'pavement', 'road', 'rock', 'sand', 'sidewalk', 'face', 'people', 'indoor', 'field', 'peak', 'wood', 'night', 'street', 'flowers', 'leaves', 'trees', 'cloudy', 'sunny', 'sunset', 'brick', 'arch', 'buildings', 'wall', 'windows', 'beach', 'high-wave', 'low-wave', 'still water', 'mirrored water', and 'snow'. These semantic concepts were also used in a previous research effort of the authors of this paper, focusing on semantic classification of personal photographs [37].

Support Vector Machines (SVMs) were adopted for classification purposes, using a Radial Basis Function (RBF) as kernel. This type of SVMs is widely used for semantic concept detection [50]. A total of 1597 images were used to train our SVM concept classifiers. The training images were retrieved from two well-known photo sets: the MPEG-7 Visual Core Experiment 2 (VCE-2) photo collection and the Corel photo collection. In addition, training was done using the low-level visual features described in Section 4.4.

To gain insight into the effectiveness of our semantic concept detectors and their semantic coverage, three human subjects created a ground truth for the 100 video clips randomly selected from the unified reference video database (see Section 4.1), using the aforementioned vocabulary of 32 semantic concepts. Next, we computed the average precision for each video clip by dividing the number of true positives by the total number of true and false positives. This allows, in its turn, computing the MAP of the 32 semantic concept detectors by calculating the mean of the average precision obtained for each video clip. We observed that the MAP of the 32 semantic concept detectors is 0.52. In addition, we observed that each semantic concept appears six times on average in a video clip. Note that the MAP of each semantic concept detector can be found in Fig. 5, computed over all 100 video clips.

4.4. Extraction of low-level visual features

The MPEG-7 color and texture descriptors, extracted using the MPEG-7 reference software [44], were used to characterize keyframes [42,43]. Specifically, the 256-D color structure (CS), 18-D color layout (CL), and 256-D scalable color (SC) descriptors were used to represent color features, whereas the 62-D homogeneous texture (HT) and 80-D edge histogram (EH) descriptors were used to represent texture features. The total dimension of the combined MPEG-7 visual descriptors is 672. The MPEG-7

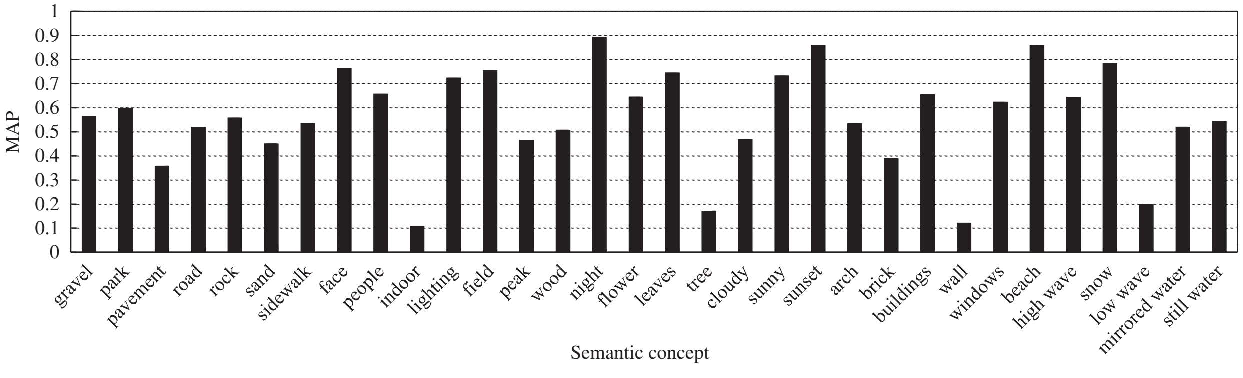

Fig. 5. MAP of the 32 semantic concept detectors.

reference software was also used to measure the L_1 distance between the extracted visual features.

Note that the image content could also have been characterized by means of SIFT [27] and bag-of-visual-words (BoVW) [51], approaches that have recently attracted a considerable amount of research interest. However, we decided to focus on the use of global MPEG-7 descriptors in our NDVC detection method because of the following three reasons: (1) the standardized MPEG-7 descriptors allow for interoperability; (2) the extraction of MPEG-7 descriptions comes with a low computational complexity; and (3) most of the semantic concepts used describe background information, which can be easily captured using color and texture features (local interest points are better suited for characterizing foreground information). Note that our NDVC detection method is compared to state-of-the-art NDVC detection methods using SIFT features and BoVW in Section 5.2.

5. Experimental results

This section reports our results obtained for five sets of experiments. The first set of experiments investigates the effectiveness of combining visual and semantic features for the purpose of NDVC detection. The second set of experiments compares the effectiveness of our NDVC detection method with the effectiveness of three other NDVC detection methods. Our third set of experiments studies the robustness of the proposed NDVC detection method against variations in semantic coverage. The fourth set of experiments looks into the influence of the effectiveness of semantic concept detection on the effectiveness of NDVC detection. Finally, the fifth set of experiments analyzes the computational complexity of our NDVC detection method and the three NDVC detection methods also used in the second set of experiments.

5.1. *Experiment I: effectiveness of combining visual and semantic features*

In this section, we assess the effectiveness of bimodal fusion of visual and semantic features for the purpose of NDVC detection, compared to the separate use of either visual or semantic features. Specifically, using the NDCR measure defined in (7), Fig. 6 compares the effectiveness of the proposed NDVC detection method with the effectiveness of an NDVC detection method only using MPEG-7 visual features and an NDVC detection method only using semantic features. For the method only making use of MPEG-7 visual features, we set $H(\mathbf{U}^q)$ to zero when computing the dissimilarity between two video clips (see (6)). On the other hand, for the method only using semantic features, we set $H(\mathbf{U}^q)$ to one. Fig. 6 shows that the proposed method considerably outperforms the method only using MPEG-7 visual features and the method only using semantic features, for all transformations considered. This demonstrates that the complementary effect created by fusing visual and semantic features has a positive impact on the effectiveness of NDVC detection.

Using the P_{miss} and R_{FA} measures defined in (8), Table 3 compares the effectiveness of the proposed NDVC detection method with the effectiveness of the method only using MPEG-7 visual features and the method only using semantic features.

We can observe that the false alarm rate of the method only using semantic features is worse than the false alarm rate of our approach. This can primarily be attributed to the fact that the discriminative power of semantic features is low when the temporal variation of semantic features is low, as discussed in Section 3.2. In contrast, the method using bimodal fusion has been designed to balance the influence of visual features and semantic features (by considering the temporal variation of the semantic features). Hence, the approach using bimodal fusion is able to achieve false alarm rates that are lower than the false alarm rates of the method only using semantic features. In addition, for the method only using MPEG-7 visual features, we can observe that the missed

Table 3
Comparison of NDVC detection effectiveness in terms of P_{miss} and R_{FA}.

	MPEG-7 visual features	Semantic features	Proposed bimodal fusion
P_{miss}	0.2723	0.0204	0.0331
R_{FA}	0.0188	0.0596	0.0235

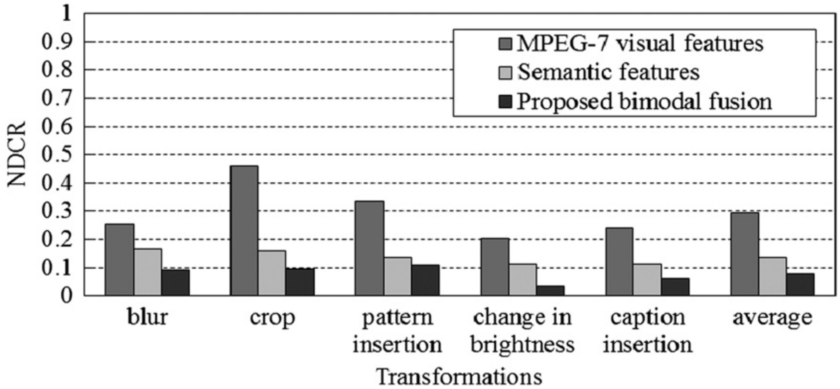

Fig. 6. Comparison of NDVC detection effectiveness in terms of NDCR.

detection probability is relatively high. This is mainly because visual features are more sensitive to transformations, compared to semantic features.

In summary, based on the experimental results shown in Fig. 6 and Table 3, we can conclude that bimodal fusion of MPEG-7 visual features and semantic features is able to facilitate robust NDVC detection.

5.2. Experiment II: comparison of NDVC detection effectiveness

In this section, we compare the effectiveness of the proposed NDVC detection method with the effectiveness of three state-of-the-art NDVC detection methods either making use of temporal ordinal measurement [23], PCA-SIFT features [27], or BoVW [51]. Both the approach relying on PCA-SIFT features and BoVW detect keypoints using Difference of Gaussians. Compared to the use of conventional SIFT features, the use of PCA-SIFT features allows for a more compact representation of the image content. In particular, the dimension of a PCA-SIFT descriptor is 36, whereas the dimension of a conventional SIFT descriptor is 128. The BoVW approach quantizes keypoints of a set of training images into different visual words, constituting a visual vocabulary. By mapping each keypoint in a test image to the nearest word, a BoVW is created, taking the form of a histogram that describes the visual content of the test image. Our experiments with the BoVW approach followed the recommendations made in [51]: the total number of visual words is 20,000 and SIFT is used to describe the keypoints of keyframes. To measure closeness between two keyframes, the *cosine similarity* was used. In addition, the frame in the middle of each shot is used as a keyframe by both the method relying on PCA-SIFT and the method making use of BoVW.

Using NDCR, P_{miss}, R_{FA}, and the ST1 score, Fig. 7 compares the mean effectiveness (averaged over ten random runs) of the proposed NDVC detection method with the mean effectiveness of the aforementioned three NDVC detection methods, considering the five transformations described in the experimental setup. In addition, to show the stability of the mean effectiveness, we also provide information about the corresponding standard deviation.

As shown in Fig. 7, the proposed NDVC detection method in general outperforms the other three NDVC detection methods. In particular, the proposed method outperforms the other three methods for the following three transformations: 'crop', 'pattern insertion', and 'caption insertion'. This can be primarily attributed to the fact that the three aforementioned transformations tend to preserve the semantic information conveyed by the original video content, whereas the NDVC detection methods only using visual features are sensitive to the aforementioned transformations. As for the transformations 'blur' and 'change in brightness', the effectiveness of the NDVC detection methods relying on either PCA-SIFT or BoVW deteriorates, compared to the NDVC detection method using temporal ordinal measurement and the proposed NDVC detection method. Temporal ordinal measurement is not affected by transformations such as 'blur' and 'change in brightness'. In addition, the NDVC detection method using PCA-SIFT is not robust against 'blur' because PCA-SIFT features make use of texture information (i.e., histogram of gradient orientations). On the other hand, for the transformations 'crop', 'pattern insertion', and 'caption insertion', the NDVC detection method using temporal ordinal measurement is not able to achieve a feasible effectiveness, whereas the effectiveness of the proposed method can be considered reliable. Fig. 7(d) also shows that the ST1 score of the proposed method is about 0.97, which is higher than the ST1 score obtained for the three other NDVC detection methods.[2] The ST1 score of 0.97 is due to a false positive, caused by the use of a strongly transformed news video clip as a query video clip. As a result, the query video clip matched with a news video clip in the reference video database different from the news video clip used to create the query video clip.

In summary, based on the experimental results shown in Fig. 7, we can conclude that the effectiveness of the proposed method is on par with or better than the effectiveness of several methods recently described in the scientific literature.

5.3. Experiment III: robustness against variations in semantic coverage

In practical NDVC systems, the number of semantic concepts used may vary from application to application. Therefore, in this section, we investigate how the number of semantic concepts used (i.e., the semantic coverage) influences the effectiveness of the proposed NDVC detection method.

Fig. 8 plots the effectiveness of NDVC detection as a function of the number of semantic concepts used. Testing was done by starting with the complete vocabulary of 32 semantic concepts (see Section 4.3), and by randomly removing two semantic concepts in subsequent iterations. When only making use of semantic features, we set $H(\mathbf{U}^q)$ to one in (6) (i.e., matching does not take into account visual features). Similarly, when only making use of MPEG-7 visual features, we set $H(\mathbf{U}^q)$ to zero.

Fig. 8 shows that the effectiveness of NDVC detection is low when the number of semantic concepts used is small, for the NDVC detection method only using semantic features. This observation holds especially true when the number of semantic concepts used ranges from two to ten. We can also observe that the effectiveness of NDVC detection increases when the number of semantic concepts used increases. Further, Fig. 8 makes clear that the NDVC detection effectiveness of bimodal fusion is highly robust against variations in the number of semantic concepts used. This can be mainly attributed to the fact that the proposed NDVC detection method is able to compensate a loss in discriminative power of the semantic features by increasing the importance of the visual features.

[2] When only making use of the MUSCLE-VCD-2007 video set, the proposed NDVC detection method was able to achieve a perfect ST1 score of 1.0.

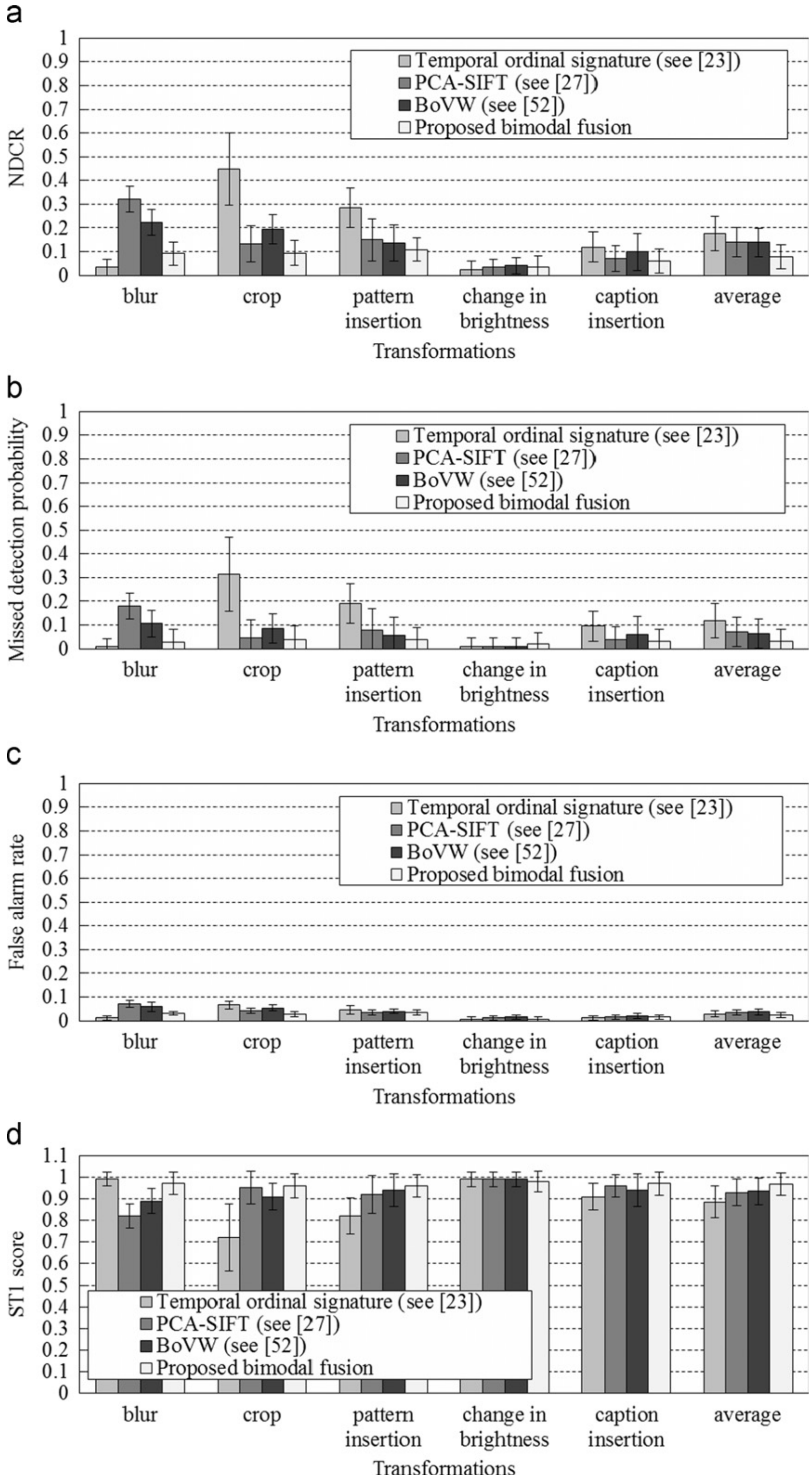

Fig. 7. Comparison of mean NDVC detection effectiveness and corresponding standard deviations for five different transformations: (a) NDCR, (b) P_{miss}, (c) R_{FA}, and (d) ST1.

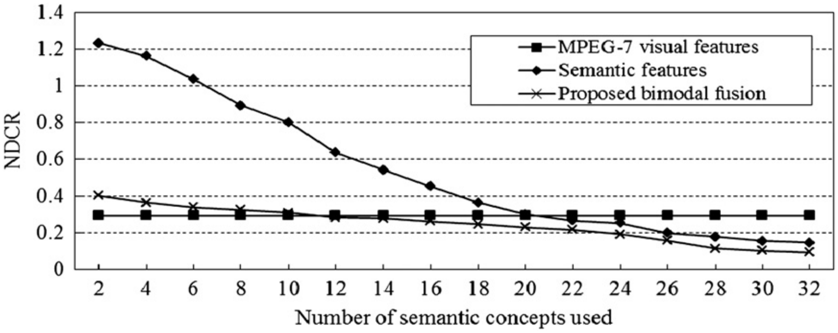

Fig. 8. NDVC detection effectiveness as a function of the number of semantics concepts used.

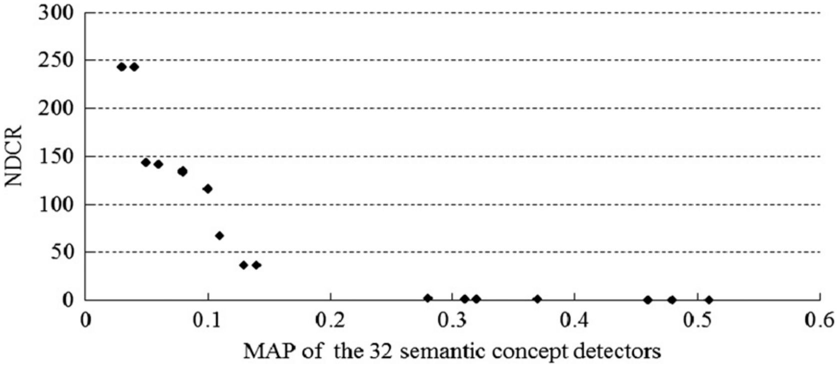

Fig. 9. Influence of the effectiveness of semantic concept detection on the effectiveness of NDVC detection.

5.4. Experiment IV: influence of effectiveness of semantic concept detection

As indicated in the introduction of this paper, semantic concept detection based on machine learning is not perfect [14]. Notwithstanding this limitation, semantic concept detection based on machine learning can still be meaningfully used. For example, [13] reports that an MAP of 0.25 is generally accepted to be sufficient for interactive search. As discussed in Section 4.3, the MAP of the 32 semantic concept detectors used in our experiments is 0.52.

To investigate the influence of the effectiveness of semantic concept detection on the effectiveness of NDVC detection, we conducted an experiment using the 100 video clips described in Section 4.1, also making use of their corresponding ground truth (see Section 4.3). Fig. 9 summarizes our experimental results. We varied the MAP of the 32 semantic concept detectors by varying the value of the threshold ζ_k (for reasons of simplicity, we used the same threshold value for all 32 semantic concept detectors in this experiment). The threshold ζ_k determines whether a corresponding concept c_k is present or absent in a video shot (see Section 3.1). Given a particular MAP value for the 32 semantic concept detectors, the corresponding data point in Fig. 9 shows the NDCR score obtained for the 100 query video clips.

As demonstrated by Fig. 9, the influence of the effectiveness of semantic concept detection on the effectiveness of NDVC detection is limited, as long as the MAP of the 32 semantic concept detectors is higher than 0.3. This result is in line with the observation made in [13] that an MAP of 0.25 is generally accepted to be sufficient for interactive search.

5.5. Experiment V: computational complexity

This section investigates the time and storage complexity of the proposed method for NDVC detection, as well as for the three NDVC detection methods described in Section 5.2. NDVC detection systems typically consist of an offline and an online part. In the offline part, an NDVC detection system generates video signatures for the set of reference video clips. In the online part, an NDVC detection system determines whether a query video clip is an NDVC of one of the reference video clips. Online processing typically consists of creating a video signature for the query video clip and video matching. This can for instance be done when uploading the query video clip to an online video repository.

To study the time complexity of the different NDVC detection methods under consideration, we measured the time needed to generate a video signature as a function of the number of shots in a query video clip. We also measured the time needed to match the 500 query video clips described in Section 4.1 as a function of the number of shots in the reference video database, relying on the 311 reference video clips also described in Section 4.1. The time needed to create a video signature includes shot segmentation, keyframe selection, and feature extraction.

The time needed to match video signatures includes matching both the visual and semantic features using a sliding window approach, as well as computing the temporal entropy for the proposed NDVC detection method. All experiments were conducted on a PC with an Intel Pentium IV 2.4 GHz CPU processor and 2 GB of system memory, running Windows XP with a 500 GB 7200 rpm hard disk.

Table 4 shows that the time complexity of creating a video signature using the proposed NDVC detection method is competitive with the time complexity of the other three NDVC detection methods studied, whereas Table 5 demonstrates that the proposed NDVC detection method allows significantly faster matching of video signatures. Further, Tables 4 and 5 illustrate that, for all NDVC detection methods used, the time complexity of both video signature creation and video signature matching is linearly dependent on the number of video shots in the query video clip and the reference video database, respectively. In other words, the time complexity of both video signature creation and video signature matching can be characterized as O(*N*), where *N* is the number of video shots processed.

Given the observation that video matching needs to be done online, having a linear complexity for the matching process may not be feasible in the context of large-scale video collections. Compared to the sliding window approach used for all NDVC detection methods investigated, more efficient matching could for instance be achieved by making use of dynamic programming [6] or hierarchical matching [28]. Further, matching can also be accelerated by indexing the MPEG-7 visual features using *kd*-trees [32] or by simply making use of parallelism, distributing a query over different nodes. As the research presented in this paper mainly focuses on proposing a novel video signature and evaluating its effectiveness, the authors would like to refer the interested reader to [6,28,32] for a discussion of more advanced techniques for optimized video indexing and matching.

Table 4
Time complexity of creating a video signature (in seconds).

Number of video shots	Temporal ordinal measurement	PCA-SIFT	BoVW	Proposed bimodal fusion
100	6.12	98.1	1391.0	81.3
1000	101.68	1071.5	1468.2	893.9
10000	992.2	10456.4	14327.1	8722.8

Table 5
Time complexity of matching video signatures (in seconds).

Number of video shots	Temporal ordinal measurement	PCA-SIFT	BoVW	Proposed bimodal fusion
100	700.21	501.3	149.2	0.81
1000	7102.17	5198.4	1686.4	8.78
10000	69422.2	51079.3	14696.1	84.19

Table 6
Storage complexity (in kbytes).

Number of video shots	Temporal ordinal measurement	PCA-SIFT	BoVW	Proposed bimodal fusion
100	0.2245	0.105	0.456	0.041
1000	2.117	1.057	4.588	0.411
10000	22.654	10.452	45.711	4.105

Besides the time complexity, we also analyzed the storage complexity of the different NDVC detection methods under consideration. Table 6 shows the amount of storage needed for different types of video signatures as a function of the number of video shots in a query video clip, illustrating that the storage complexity of the different types of video signatures is linearly dependent on the number of video shots. In other words, the storage complexity can be characterized as O(*N*), where *N* is the number of video shots processed.

Note that, whereas storing the MPEG-7 descriptors needs about 0.4 kbytes per shot, storing the semantic features only requires 4 bytes per shot when a vocabulary is used of 32 semantic concepts (given that the proposed semantic video signature consists of a sparse matrix of binary values). It is possible to achieve a more compact representation for the semantic video signature by making use of run-level coding followed by variable length coding, or by making use of more advanced techniques regarding the representation and processing of sparse matrixes (see for instance [52,53]).

6. Conclusions and directions for future research

This paper aimed at furthering the understanding of the use of a semantic approach towards the task of NDVC detection. Specifically, this paper introduced a novel NDVC detection method that makes use of bimodal fusion of low-level visual features and high-level semantic features in order to facilitate robust identification of NDVCs. We construct a semantic video signature by using trained classifiers to detect the presence and absence of 32 semantic concepts along the temporal axis of a video clip. The discriminative power of this semantic video signature depends on the temporal variation of the semantic concepts: a high temporal variation of the semantic concepts denotes a high discriminative power of the semantic modality and vice versa. When detecting that the temporal variation of the semantic concepts is low, we make use of MPEG-7 visual features to compensate for the low discriminative power of the semantic features. We quantify the temporal variation of the semantic concepts, and thus their discriminative power, by computing the temporal entropy of a semantic video signature. We also make use of temporal entropy to automatically balance the influence of visual and semantic features on the NDVC detection process.

Experimental results obtained for the MUSCLE-VCD-2007 and the TRECVID 2008 video sets demonstrate that bimodal fusion of visual and semantic features is able to identify NDVCs with a low missed detection rate and a low false alarm rate. In addition, for the task of NDVC detection, the combined use of visual and semantic features yields a better effectiveness than the separate use of either visual or semantic features. Further, we showed that the effectiveness of the proposed NDVC detection method is on par with or better than the effectiveness of three state-of-the-art NDVC detection methods either making use of temporal ordinal measurement, PCA-SIFT features, or BoVW. We also demonstrated that the influence of the effectiveness of semantic concept detection on the effectiveness of NDVC detection is limited, as long as the mean average precision (MAP) of the semantic concept detectors is higher than 0.3. Finally, we illustrated that the computational complexity of our NDVC detection method is competitive with the computational complexity of the three aforementioned NDVC detection methods.

The design and evaluation of NDVC detection systems is a well-studied topic, touching upon many aspects of computer science and engineering. As such, several directions for future research can be identified. In order to allow for a more extensive semantic concept vocabulary, we plan to investigate the use of a model-free approach towards video concept detection. In addition, we plan to pay attention to a more compact representation of the proposed video signature, as well as optimized indexing and matching techniques. Further, we also plan to look into partial matching.

Acknowledgments

This research was supported by the Basic Science Research Program of the National Research Foundation (NRF) of Korea, funded by the Ministry of Education, Science and Technology (research grant: 2011-0011383).

References

[1] R.D. Oliveira, M. Cherubini, N. Oliver, Looking at near-duplicate videos from a human-centric perspective, ACM Transactions on Multimedia Computing, Communications, and Applications 6 (3) (2010).

[2] X. Wu, C.-W. Ngo, A. Hauptmann, H.-K. Tan, Real-time near-duplicate elimination for web-video search with content and context, IEEE Transactions on Multimedia 11 (2) (2009) 196–207.

[3] A. Joly, O. Buisson, C. Frelicot, Content-based copy retrieval using distortion-based probabilistic similarity search, IEEE Transactions on Multimedia 9 (2) (2007) 293–306.

[4] C. Chih-Yi, C. Chu-Song, C. Lee-Feng, A framework for handling spatiotemporal variations in video copy detection, IEEE Transactions on Circuits and Systems for Video Technology 18 (3) (2008) 412–417.

[5] M. Bober, P. Brasnett, S. Paschalakis, Recent developments on standardisation of MPEG-7 Visual Signature Tools, in: Proceedings of the IEEE International Conference on Multimedia & Expo, 2010, pp. 1347–1352.

[6] X.S. Hua, X. Chen, H.J. Zhang, Robust video signature based on ordinal measure, in: Proceedings of ICIP, 2004, pp. 685–688.

[7] C. Kim, B. Vasudev, Spatiotemporal sequence matching for efficient video copy detection, IEEE Transactions on Circuits and Systems for Video Technology 15 (1) (2005) 127–131.

[8] C. Cotsaces, N. Nikolaidis, I. Pitas, Semantic video fingerprinting and retrieval using face information, Signal Processing: Image Communication 24 (7) (2009) 598–613.

[9] J. Law-To, L. Chen, A. Joly, I. Laptev, O. Buisson, V. Gouet-Brunet, N. Boujemaa, F. Stentiford, Video copy detection: a comparative study, in: Proceedings of ACM CIVR, 2007, pp. 371–378.

[10] G. Carneiro, A.B. Chan, P.J. Moreno, N. Vasconcelos, Supervised learning of semantic classes for image annotation and retrieval, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (3) (2007) 394–410.

[11] O. Chapelle, P. Haffner, V.N. Vapnik, Support vector machines for histogram-based image classification, IEEE Transactions on Neural Networks 10 (5) (1999) 1055–1064.

[12] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, H. Wactlar, Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news, IEEE Transactions on Multimedia 9 (5) (2007) 958–966.

[13] C.G.M. Snoek, A.W.M. Smeulders, Visual-concept search solved? IEEE Computer 43 (6) (2010) 76–78

[14] J. Yang, A.G. Hauptmann, (Un)Reliability of video concept detection, in: Proceedings of the International Conference on Content-based Image and Video Retrieval, 2008, pp. 85–94.

[15] P. Over, W. Kraaij, A.F. Smeaton, TRECVID 2008—an overview, in: Proceedings of the TRECVID 2008, NIST, USA, 2008.

[16] J. Law-To, A. Joly, N. Boujemaa, MUSCLE-VCD-2007: a live benchmark for video copy detection, 2007, available on 〈http://www-rocq.inria.fr/imedia/cive-bench〉.

[17] H.S. Min, J.Y. Choi, W. De Neve, Y.M. Ro, Near-duplicate video detection using temporal patterns of semantic concepts, in: Proceedings of the IEEE International Symposium on Multimedia, 2009, pp. 1–8.

[18] H.S. Min, W. De Neve, Y.M. Ro, Towards using semantic features for near-duplicate video detection, in: Proceedings of the IEEE International Conference on Multimedia & Expo, 2010, pp. 1364–1369.

[19] A. Basharat, Y. Zhai, M. Shan, Content based video matching using spatiotemporal volumes, Journal of Computer Video and Image Understanding 110 (3) (2008) 360–377.

[20] O. Chum, J. Philbin, A. Zisserman, Near duplicate image detection: min-hash and tf-idf weighting, in: Proceedings of the British Machine Vision Conference, 2008, pp. 73–80.

[21] S. Poullot, M. Crucianu, O. Buisson, Scalable mining of large video databases using copy detection, in: Proceedings of the ACM International Conference on Multimedia, 2008, pp. 61–70.

[22] M.-C. Yeh, K.-T. Cheng, A compact, effective descriptor for video copy detection, in: Proceedings of the ACM International Conference on Multimedia, 2009, pp. 633–636.

[23] L. Chen, F.W.M. Stentiford, Video sequence matching based on temporal ordinal measurement, Pattern Recognition Letters 29 (13) (2008) 1824–1831.

[24] T.C. Hoad, J. Zobel, Detection of video sequences using compact signatures, ACM Transactions on Information Systems 24 (1) (2006) 1–50.

[25] G. Leon, H. Kalva, B. Furht, Video identification using video tomography, in: Proceedings of the IEEE International Conference on Multimedia & Expo, 2008, pp. 1030–1033.

[26] P.-H. Wu, T. Thaipanich, C. J. Kuo, Detecting duplicate video based on camera transitional behavior, in: Proceedings of the ICIP, 2009, pp. 237–240.

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

[28] X. Wu, A. G. Hauptmann, C.-W. Ngo, Practical elimination of near-duplicates from web video search, in: Proceedings of the the ACM International Conference on Multimedia, 2007, pp. 218–227.

[29] J. Zhu, C.H. Steven, R. Michael, S. Yan, Near-duplicate keyframe retrieval by nonrigid image matching, in: Proceedings of the ACM International Conference on Multimedia, 2008, pp. 41–50.

[30] J. Law-To, G.-B. Valerie, B. Olivier, B. Nozha, Local behaviours labelling for content based video copy detection, in: Proceedings of the ICPR, 2006, pp. 232–235.

[31] X. Zhou, Lei Chen, A. Bouguettaya, N. Xiao, J.A. Taylor, An efficient near-duplicate video shot detection method using shot-based interest points, IEEE Transactions on Multimedia 11 (5) (2009) 879–891.

[32] O. Küçüktunç, M. Baştan, U. Güdükbay, Ö. Ulusoy, Video copy detection using multiple visual cues and MPEG-7 descriptors, Journal of Visual Communication and Image Representation 21 (8) (2010) 838–849.

[33] M.-C. Yeh, K.-T. Cheng, Video copy detection by fast sequence matching, in: Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR), 2009, pp. 1–7.

[34] S. Poullot, O. Buisson, M. Crucianu, Scaling content-based copy detection to very large databases, Multimedia Tools and Applications 47 (2) (2010) 279–306.

[35] X. Wu, W.-L. Zhao, C.-W. Ngo, Near-duplicate keyframe retrieval with visual keywords and semantic context, in: Proceedings of the ACM International Conference on Image and Video Retrieval, July 2007, pp. 162–169.

[36] W. Meng, H. Xian-Sheng, H. Richang, T. Jinhui, Q. Guo-Jun, S. Yan, Unified video annotation via multigraph learning, IEEE Transactions on Circuits and Systems for Video Technology 19 (5) (2009) 733–746.

[37] S. Yang, S.K. Kim, Y.M. Ro, Semantic home photo categorization, IEEE Transactions on Circuits and Systems for Video Technology 17 (3) (2007) 324–335.

[38] X. Anguera, P. Obrador, T. Adamek, D. Marimon, N. Oliver, Telefonica research content-based copy detection TRECVID submission, in: Proceedings of the NIST TRECVID 2009 Workshop, Notebook Paper, 2009.

[39] E.-J. Hah, P. Schmutz, A.N. Tuch, D. Agotai, M. Wiedmer, K. Opwis, Cinematographic techniques in architectural animations and their effects on viewers' judgment, International Journal of Design 2 (3) (2008) 29–41.

[40] C. Petersohn, Fraunhofer HHI at TRECVID 2004: shot boundary detection system, in: TREC Video Retrieval Evaluation Online Proceedings, 2004.

[41] Guidelines for the TRECVID 2008 Evaluation. Available on: 〈http://www-nlpir.nist.gov/projects/tv2008/tv2008.html〉.

[42] T. Sikora, The MPEG-7 visual standard for content description—an overview, IEEE Transactions on Circuits and Systems for Video Technology 11 (6) (2001) 696–702.

[43] B.S. Manjunath, J.-R. Ohm, V.V. Vasudevan, A. Yamada, Color and texture descriptors, IEEE Transactions on Circuits and Systems for Video Technology 11 (6) (2001) 703–715.

[44] MPEG-7 XM software. Available on: 〈http://www.lis.ei.tum.de/research/bv/topics/mmdb/empeg7.html〉.

[45] Video Structure Detection—Temporal Video Segmentation. Available on: 〈http://www.hhi.fraunhofer.de/en/departments/interactive-media-human-factors/overview/video-structure-detection/〉.

[46] Description of the MUSCLE-VCD-2007 video database. Available on: 〈http://www-rocq.inria.fr/imedia/civr-bench/VideoDatabase.html〉.

[47] W. Kraaij, P. Over, J. Fiscus, A. Joly, Final CBCD evaluation plan TRECVID 2008 (v1.3), June 2008.

[48] Final list of video transformation. Available on: 〈http://www-nlpir.nist.gov/projects/tv2008/final.cbcd.video.transformations.pdf〉.

[49] CBCD Evaluation Plan TRECVID 2010, 2010. Available on: 〈http://www-nlpir.nist.gov/projects/tv2010/Evaluation-cbcd-v1.3.htm#eval〉.

[50] K.-R. Müller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, An introduction to kernel-based learning algorithms, IEEE Transactions on Neural Networks 12 (2) (2001) 181–201.

[51] W.-L. Zhao, X. Wu, C.-W. Ngo, On the annotation of web videos by efficient near-duplicate search, IEEE Transactions on Multimedia 12 (5) (2010) 448–461.

[52] J. Willcock, A. Lumsdaine, Accelerating sparse matrix computations via data compression, in: Proceedings of the International Conference on Supercomputing, 2006, pp. 307–316.

[53] J.R. Gilbert, C. Moler, R. Schreiber, Sparse matrices in MATLAB: design and implementation, SIAM Journal on Matrix Analysis and Applications 13 (1) (1992) 333–356.